MACHINE LEARNING ENHANCEMENTS FOR KNOWLEDGE DISCOVERY IN MINERAL EXPLORATION AND IMPROVED MINERAL RESOURCE CLASSIFICATION

by

Selcuk Ilkay Cevik

A thesis submitted to the Department of Department of Mining Engineering

In conformity with the requirements for

the degree of Master of Applied Science

Queen's University

Kingston, Ontario, Canada

(September 2020)

To my wife, Dilan.
To my family, Fatma, Muhittin and Berkay.

Abstract

Contemporary issues of the mining industry, such as declining grades, increasing depth of deposits, cost per discoveries and demand for the raw materials are driving the mining industry to develop and adopt improved technologies. The main objective is to achieve success in exploring deeper parts of the earth and improve mining and processing technologies. These emerging technologies generate enormous amounts of data, therefore utilizing all the available data efficiently has become one of the biggest challenges for geoscientists and engineers.

Machine learning (ML) algorithms are being deployed to handle the abundance of data in many traditional fields (e.g., computer vision, natural language processing and genetics) due to their fast processing capability. Despite its potential benefits, ML has not been fully adopted by the mineral exploration and resources sector yet. This thesis demonstrates that ML can assist in improving mineral exploration strategies and mining procedures by enhancing the understanding of processes from high dimensional data and by automating operations. An overview of state-of-the-art applications of different ML approaches and their relevance to the mineral exploration and resources sectors is presented by means of two original case studies. Furthermore, a generic workflow for ML application in geoscience is proposed to highlight good practices in their implementation.

The first case study encompasses an example of a multivariate analysis applied to a lithogeochemical dataset from the Vazante-Paracatu District, Brazil, in order to provide insights about processes related to the base metal mineralizing system. The complex relationships between the data and the mineral occurrences are revealed to assist in finding new targets for zinc exploration.

The second case study is an example of ML applied to resource classification, which normally relies on expert assessment of a qualified person to determine if the blocks of a 3D mineral resource model are classified as measured, indicated, or inferred. This study reveals that ML can assist to increase time efficiency of the task and improve consistency by automating the process.

These applications demonstrate the relevance of ML methods in supporting knowledge discovery in geosciences and engineering and automating processes for improved consistency in the results.

Co-Authorship

The two case studies presented in Chapter 3 are prepared as standalone manuscripts for peer-reviewed journals.

The first case study is titled as "*A combined multivariate approach analyzing geochemical data for knowledge discovery: The Vazante – Paracatu Zinc District, Minas Gerais, Brazil*", co-authored by the two supervisors of the candidate, Dr. Julian M. Ortiz and Dr. Gema R. Olivo, and submitted to the "*Journal of Geochemical Exploration"*. This paper is under review at the time of this thesis submission.

The second case study is titled as *"On the use of machine learning for mineral resource classification"* is co-authored by Dr. Julian M. Ortiz, supervisor of the candidate, Dr. Oy Leuangthong and Dr. Antoine Caté, from SRK Toronto office, and will be submitted to a peer-reviewed journal.

Acknowledgements

During my years at Queen's, I had experienced an extraordinary feeling of self-development that I would never envision. For this, I would like to thank my supervisors, Dr. Julian Ortiz and Dr. Gema Olivo. Dr. Ortiz created an ever-lasting enthusiasm for my research area that pushed me to extend the limits of my curiosity. Dr. Olivo provided me a unique mindset of exploration geology that enabled me to pursue research in this area.

I was privileged to work with my colleagues in SRK Toronto during the research that provided me with a rare opportunity of discussing practical aspects of my research area for which I thank them all. I would especially like to thank Dr. Oy Leuangthong and Dr. Antoine Caté. This research is benefited greatly from their ideas and invaluable time.

Predictive Geometallurgy and Geostatistics Group and The Robert M. Buchan Department of Mining at Queen's University provided a unique, multicultural research environment. I especially thank my fellow researchers, Álvaro Riquelme, Sebastián Ávalos, William Midkiff and Ali Hashemi for their friendship.

I thank Efdal Ölçer and Ali Can Akpınar, two inspirational people I know since my early career years. Their guidance shaped my personal and professional attitude.

I am most grateful to the Ministry of National Education of the Republic of Turkey as it would not be possible to pursue my degree without their financial support. I also greatly appreciate the financial supports of MITACS Accelerate, SRK Toronto, and NSERC. I thank Nexa Resources, for allowing data collection, Colin Aldis, and Neil Fernandes for sharing their data and knowledge on the Vazante - Paracatu District in Brazil.

I thank my family, Muhittin, Fatma, and Berkay for their endless support during my career even if it means that I am living far from them for a long period. Lastly and most importantly, I

thank my wife, Dilan, for her unconditional sacrifices during these years. Without her love and support, it would not be possible for me to pursue this path.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

The mineral resources sector needs to transform to address several contemporary issues such as challenging financial, environmental and social conditions, declining grades, increasing depth of deposits (Ortiz, 2019; Douce, 2016: Arndt et al., 2017) and the increasing costs per discovery (Groves and Santosh, 2015). Besides, there is an ever-increasing demand for raw materials (Lishcuk, 2019 and references therein). This drives the mining industry to focus on deposits that were not previously considered as economical due to several reasons such as depth of formation, geological and mineralogical complexity, low grades, fine-grained ore textures, or presence of deleterious elements. Improvements in sensing technologies make it possible to explore deeper parts of the Earth, while developments in the mining, mineral processing and extracting methods make it possible to exploit low-grade masses up to 2-3 km depth and give rise to the development of more advanced systems throughout the mining value chain which requires the collaboration of many professionals (Arndt et al., 2017; Cate et al., 2017; Lishcuk, 2019). These developments give opportunity to geometallurgy to become an important approach that attempts to integrate whole mining systems to optimize the entire process which eventually facilitates the use of natural resources in the most sustainable way possible (Lishcuk, 2019; Arndt et al., 2017).

The resolution of the available data is a controlling factor to optimize mining and metallurgical processes and exploration strategies. Hence, many contemporary technologies, e.g. improved sensing technologies and increased computational resources, have focused on generating enormous amounts of high-quality data for decades, which results in immense developments and gave rise to big data. In addition to this, the democratization of data through cloud systems, increasing the open-source nature of the information, and applications enhanced the accessibility of these developments, and therefore, increased the rate of improvements even further by allowing

many individuals to conduct research on large datasets. These developments evolved geoscience from a data-poor field to a data-rich field (Karpatne et al., 2018). As a result, managing and utilizing all the available geoscience data to get as much useful information as possible and gain insights about the processes have become one of the biggest challenges for geoscientists.

Machine learning (ML) algorithms are able to construct an approximated function for an underlying process that generates patterns or regularities in the data (Alpaydin, 2020) and use the data to improve performance in terms of a given metric (Mitchel, 1997). Because of their fast processing capability, these algorithms are being deployed to handle the big data in many traditional fields such as computer vision, natural language processing and genetics. They are able to make data-driven inferences and recognize complex relationships by allowing exploration of large function spaces. In addition to these capabilities, the availability of powerful and easy to use ML tools have led to an increasing interest in ML applications among geoscientists as well (Bergen et al., 2019).

Arndt et al., 2017 argue that a balance between supplies and increasing demand of the mineral resources will only be achieved by the combination of discovering new deposits, which are getting deeper as most of the near surface mineral deposits have already been discovered, and by converting existing resources to reserves that were previously considered as uneconomical. In both cases, the effective utilization of the data offered by ML methods is crucial to manage and reduce the risk associated with uncertainties, and to make better-informed decisions. ML also provides tools to automate processes currently done by experts or operators, hence it reduces the human-related errors, improves consistency and time efficiency (Ortiz et al., 2020).

Like many other fields in geoscience, challenges in the mineral resource sector partly comes from the high dimensional nature of the problem and fuzzy boundaries between concepts and definitions which can be alleviated by better understanding of the information provided in the datasets. Long-standing debates related to the formation of certain types of mineral deposits are the function of the complex processes involved in their formation and the high dimensional nature of the problem. ML can assist in improving the understanding of these processes through knowledge discovery approach.

Despite its potential benefits, ML has not been fully adopted by the mineral exploration and resources sector yet. We claim that ML can assist in improving mineral exploration strategies and mining by enhancing the understanding of processes from high dimensional data (knowledge discovery) and by automating operations. This approach has potential to optimizing consistency in the exploration and mining projects based on expert decisions. The consequence of embracing ML in mineral exploration and resources characterization is an improvement in the efficiency of the entire mining value chain as a part of an integrated geometallurgical approach. ML is also an enabler for automation of mining and metallurgical decisions, having potential to contribute significantly to the entire mine cycle, in particular if integrated thoroughly.

The aim of this thesis is to provide an overview of state-of-the-art applications of different ML approaches and present their relevance to the mineral exploration and resources sectors by means of two original case studies. A generic workflow of a good practice for ML application in geoscience is also proposed.

In the first case study, we show an example of a multivariate analysis of an existing lithogeochemical dataset of rock samples, collected throughout a basin in a metallogenic belt, to

provide insights about processes related to the mineralizing system at the Vazante-Paracatu District, Brazil. It is an example of an application where knowledge discovery is a concern. The complex relationships between the data and the mineral occurrences are revealed to identify zones with pre-enrichment and syn-ore depletions and assist in finding new targets for exploration.

The second case study is an example of ML applied to resource classification process, which normally relies on expert assessment of a qualified person to determine which blocks of a 3D mineral resource model are classified as measured, indicated, or inferred. The aim of the study is to increase time efficiency of the task and improve consistency by automating the process rather than revealing relationships.

Lastly, current challenges and gaps of ML applications in the mineral resource sector are discussed and possible research directions are indicated.

These subjects are structured as follows in this thesis: Chapter 2 provides an overview of the major ML types, outlines their relevance to challenges in the geoscience domain, and presents state of the art applications in the mineral resource sector. A generic workflow for ML applications in geoscience domain that contains good practices is also provided in this section. Chapter 3 presents the two case studies mentioned above. Chapter 4 concludes the thesis with an analysis of the implications of ML applications in the mineral resource sector as a whole and discusses the challenges of applying ML that are specific to the geoscience data and discusses future possible research directions.

Chapter 2 Literature Review

2.1. Introduction

In Chapter 1, machine learning algorithms (MLA) are briefly introduced as algorithms that are able to make approximations for an underlying process that is assumed to exist and is responsible for the patterns or regularities found in the data. Identification of those patterns in the data and generating approximation models are useful for a variety of reasons. By looking at the models and/or identified patterns, we can gain insights about a process (descriptive models for knowledge discovery) or in other instances, we use these models simply to predict an output given an input and automate a process (predictive models) (Alpaydin, 2020). Carranza (2008), defines *models* as making descriptions, representations or predictions about an indirectly observable and complex real-world system, via quantitative analysis of relevant data. As humans, we have been generating models and using them to describe or predict many things before the computer aided approaches, i.e. machine learning methods. In the field of geology, Total Alkali-Silica (TAS) scheme (Le Maitre, 1984), which is used to predict the class of a volcanic rock as a function of SiO_2 and K_2O+Na_2O (Figure 2.1A) is a good example of human generated statistical predictive methods. However, our efficiency to detect patterns decreases as the dimension of the data increases. Therefore, machine learning algorithms provide opportunities to extend these pattern recognition abilities into high dimensional, large data sets.

In this chapter, the definition of the two main types of machine learning modeling processes is presented:

- *Supervised learning*, where a desired output or label is available for a given input set of observations and the algorithm is expected to construct a function f to relate the input data to the given output; and
- *Unsupervised learning*, where the algorithm is expected to reveal patterns or regularities that are more frequent than others (Alpaydin, 2020).

Commonly used supervised and unsupervised machine learning methods are tabulated in (Table 2.1) to provide the reader with a list of tools, and their use, that are available for a geoscientist with a machine learning toolbox. Some of those methods were applied to the case studies presented in this thesis. Chapter 2 concludes with an overview of the state-of-the-art applications in mineral exploration and resource sector and a brief discussion of the challenges that could be addressed in future studies.

2.2. Supervised Learning

A supervised learning task uses a set of known examples of inputs and their associated outputs to approximate a functional relationship to predict the output from the input data. It is considered as *classification* when the desired output is a discrete value, a category, or simply a class, and as *regression* when it is a continuous value (James et al., 2013).

A supervised model uses a set of examples or observations (Alpaydin, 2020):

$$X = \{x^t, y^t\}_{t=1}^{N}$$

where t is index for one of the N observations, x^t is the input and y^t is the associated output. The input can have any arbitrary dimension. The output may be one of K classes in classification problems, or a scalar or vector in the case of regression problems.

A generalized form of MLA that will be used for approximation is expressed as follows:

$$y = f(x|\theta)$$

where f is the model, θ is its parameters, x is the input data and y is the output (ground truth). MLA aim at optimizing the parameters θ, to minimize the approximation error, which is the difference between the estimates and the ground truth. *The loss function* is the difference between expected output y^t and the model's approximation, $f(x^t|\theta)$:

$$E(\theta|X) = \sum_t L\,(y^t, f(x^t|\theta))$$

The optimization procedure to find the optimum parameters, θ^*, that minimizes the loss, can be expressed as:

$$\theta^* = argmin\, E(\theta|X)$$

where $argmin$ returns the argument that minimizes the loss function. A demonstration of the classification task is presented in the next example (Figure 2.1).

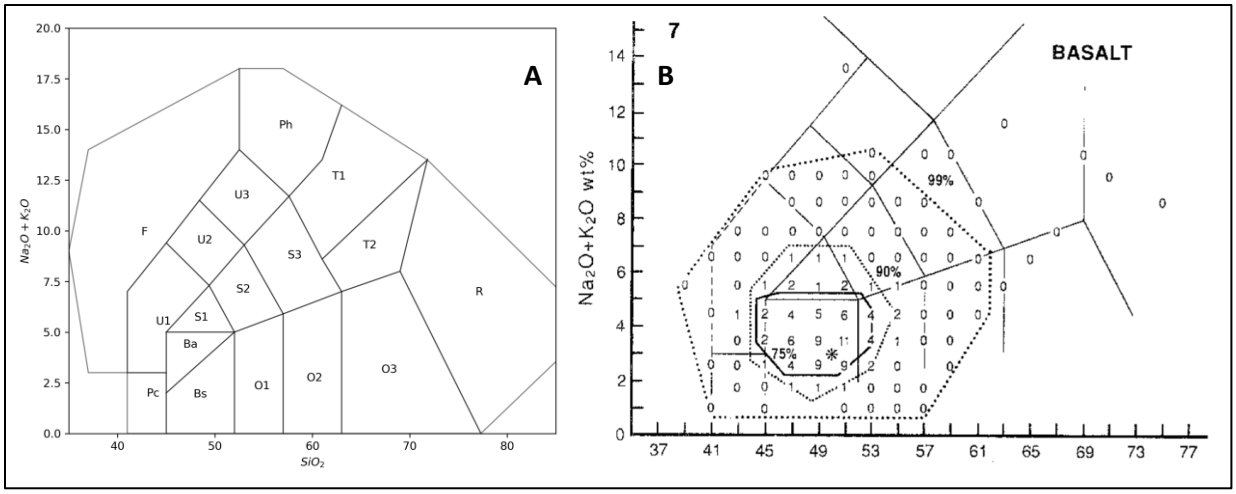


Figure 2.1 TAS classification scheme on the left, and an example of the construction of the field boundaries. The figure on the right is retrieved from Le Bas et al., 1992. Abbreviations; Pc: Picro-basalt, Bs + Ba: Basalt, O1: Basaltic andesite, O2: Andesite, O3: Dacite, R: Rhyolite, S1: Trachy-basalt, S2: Basaltic trachy-andesite, S3: Trachy-andesite, T1: Trachyte, T2: Trachydacite, U1:Tephyrite Basanite, U2:Phono-tephrite, U3:Tephri-phonolite, Ph:Phonolite, F:Foidite.

The Total Alkali-Silica (TAS) scheme is first introduced by Le Maitre (1984), to facilitate classification of volcanic rocks when mineral modal data are not available. The details of how boundaries between classes are constructed are provided in Le Bas et al. (1992). In simple terms, a large data set, namely CLAIR (Le Maitre et al., 1978), that comprises some 15,000 volcanic rock samples around the world with lithogeochemical analysis and their descriptions by petrologist or geologists, are used to construct the classification scheme. Lithogeochemistry of the rocks are used as inputs, x, namely SiO_2 in the x-axis and total alkali (K_2O+Na_2O) in the y-axis, and nomenclature by experts used as target variable, y, feed a classification scheme that approximates the boundaries between classes. The principal criteria to define the boundaries between classes is to minimize the degree of overlap between adjacent fields (Le Bas et al., 1992). Percentage frequency distribution plots (Figure 2.1B) are used to assess overlap between fields. This is a typical example of combination of data and knowledge driven classification approach manually conducted by an expert which allows some degree of subjectivity, e.g. decision of omitting some of the samples

with low confidence or reclassifying some of the classes based on local knowledge or preconceptions.

A supervised MLA performs similarly in the sense that it uses the inputs and outputs and tries to minimize the approximation error given the model parameters without imposing any prior knowledge. A simple machine learning method, namely k-nearest neighbor (kNN), that allows k number of nearest neighbors, in feature space, to vote for classification of an unknown sample, is used to construct a classification scheme similar to TAS. A synthetic dataset is created for demonstration purpose (Figure 2.2A). The class boundaries are determined with kNN method by using this data set (Figure 2.2B). Although some noticeable differences exist, e.g. Ph vs T1, T1 vs T2, etc., this machine learning approach produces class boundaries that are similar to the original TAS scheme. Also note that, a synthetic data set created to demonstrate the classification approach was used, but the original data set could yield results that are closer to the original TAS scheme. Nevertheless, this simple example shows the similarities between the two modeling approaches. Although the original TAS scheme that combines knowledge and data driven approach allows the expert to have more control over the predictive model, it is not hard to imagine that complexity of the problem will increase with the increasing feature dimension. For example, in Le Bas et al. (1992), boundaries in TAS scheme were chosen to minimize the overlaps between the fields that comprise approximately 75% of the samples from the corresponding class (Figure 2.1B). Adding another predictor besides SiO_2 and total alkali as a third axis could help to reduce this overlap, and thus improve the classification performance. However, increasing dimension would make the manual classification by the expert extremely difficult. MLA can easily integrate multiple dimensions to the approximation process and produce results that are reproducible and less subjective.

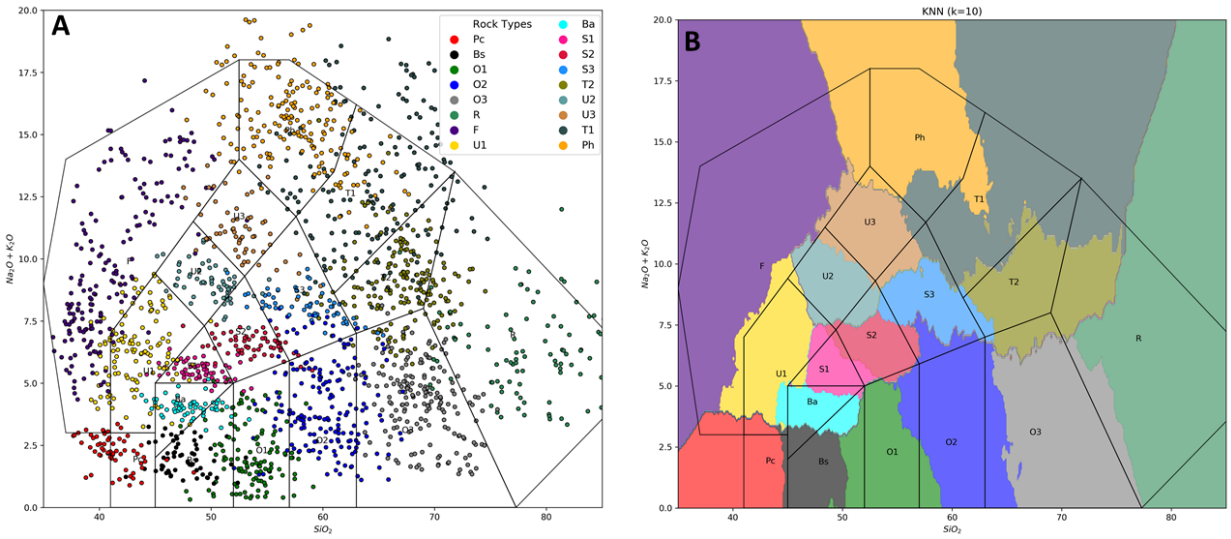


Figure 2.2 A synthetic set of samples created by a random function is on the left and ML produced class boundaries by using the synthetic data set on the right.

Obviously, the ultimate goal of such a classification scheme is to be able to classify future, unknown samples with high accuracy rather than the samples that are used in training. Allowing the algorithm to learn *too much* from the training data, may result in something called, *overfitting*, and conversely, constraining the algorithm *too much* results in *underfitting*. Overfitting can be envisioned as memorizing the data, which makes the algorithm poor in generalizing the rules that are learned during the training, and underfitting can be considered as insufficient learning. Both overfitting and underfitting are demonstrated in Figure 2.3 and Figure 2.4, respectively, with the TAS example.

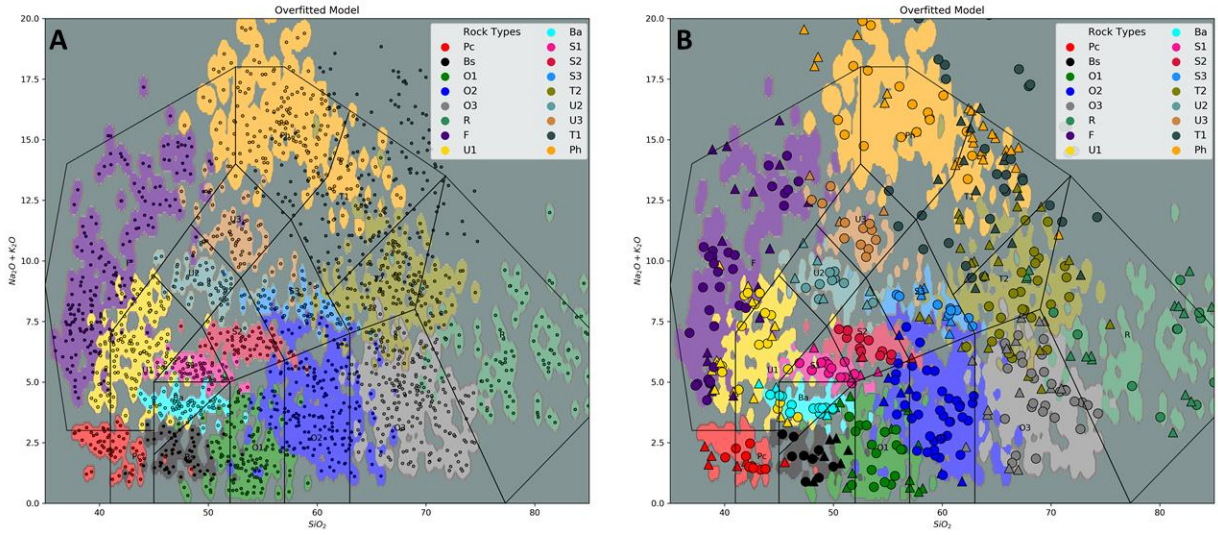


Figure 2.3 Overfitting example. The ML model overfits to the training data on the left, i.e. memorizes the training data instead of generalizing the classification rules and does not perform well on an independent validation set on the right. Triangles represent misclassified samples.

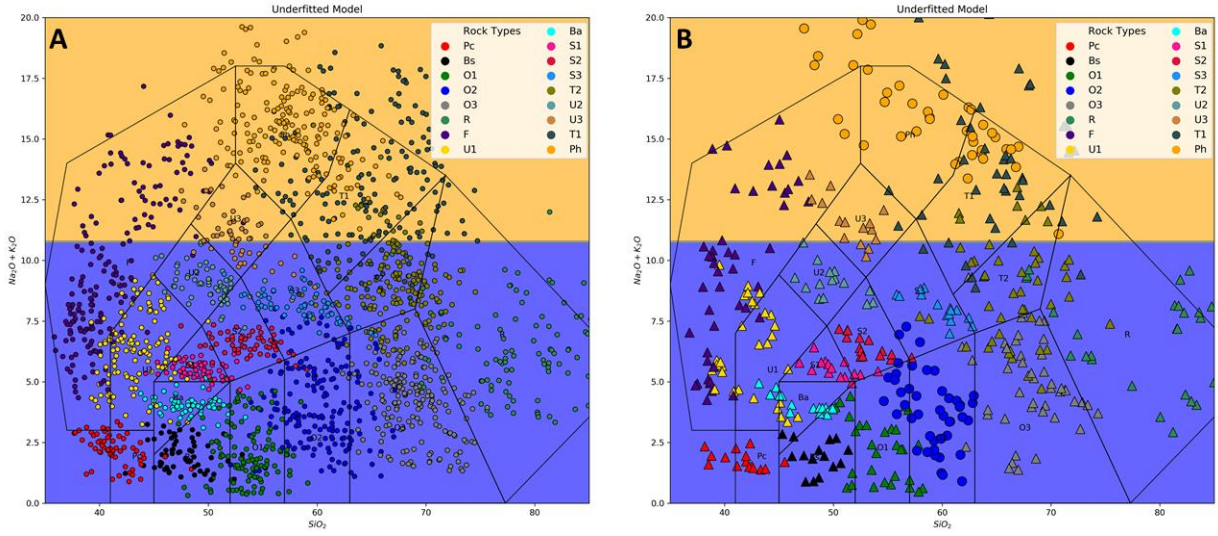


Figure 2.4 Underfitting example. The ML model underfits to the training data on the left, i.e. could not learn enough to separate classes and does not perform well on an independent validation set on the right. Triangles represent misclassified samples.

These show that it is important to test the MLA model with a group of samples that are independent from the training samples, also called *validation samples*, for its predictive capabilities and to tune the parameters to increase the model performance. A validation set is required at minimum, and a

second independent data set, called *test set* that is not used as part of training or validation sets, is required for best practice to report the predictive performance of an MLA (Alpaydin, 2020). Usually data is scarce and leaving out two independent data set is not preferable. In these cases, after leaving out the test set, data is split into K approximately equal size subsets, and one of these subsets is left out and the remaining $K-1$ parts are used to train a model and the left-out part is used for validation. Cross validated prediction error is then reported after repeating this procedure for each K part. This is called K-Fold cross-validation (Friedman et al., 2001). In most of the ML applications, splitting the data into K parts is done by random splitting after shuffling the data set. However, it should be noted that the datasets that are used in mineral exploration and resource sector usually have spatial dependencies, also called *spatial autocorrelation*, meaning that the independence requirement in cross-validation is violated when data is shuffled and split randomly because of the spatial autocorrelation. In other words, if data is not split into spatially distinct sets for cross-validation, predictive performance is overestimated because of the dependency between training and validation sets. This problem is acknowledged in the literature, and variations of the cross-validation approach that addresses the spatial dependency have been proposed (Roberts et al., 2017; Pohjankukka et al., 2017).

Another example for a supervised task in mineral exploration could be predicting the potential of having a mineral deposit at an unknown location as a function of available geoscience data such as lithogeochemical, geophysical and remote sensing features as inputs. By providing examples of known mineral occurrences, *positive samples*, and known barren zones, *negative samples*, as output labels, a model can be trained. In such an example, a geologist may be interested in knowing the predicted result of the machine learning model for the location of interest. Perhaps, it would be interesting as well to achieve *knowledge discovery*, that is, by analyzing the approximated

function, the geologist can understand what features should be expected in a location with high probability of having a mineral deposit.

On the other hand, an automation task such as recognizing possible faults for a hydrocarbon exploration study where seismic images are available as input and the location of some faults are provided as desired labels, might not require to gain an understanding about the process necessarily, instead high accuracy and time efficiency would be critical (Zhang et al., 2019; Li et al., 2019).

It is these abilities to discover complex relationships between the observations or variables and to process large amounts of data, that make the supervised learning algorithms highly attractive for researchers and practitioners who study in the mineral resource sector, where assimilation and evaluation of high dimensional, multidisciplinary data is critical to explore and exploit mineral resources effectively.

2.3. Unsupervised Learning

Unsupervised learning methods are different than supervised ones, in that there is no response variable, y^t, and the aim is not to make predictions, but to conduct an exploratory study to recognize the global and local structures in the data. The data is made of N observation with arbitrary dimensional input, $X = \{x^t\}_{t=1}^N$. In other words, the aim is to reveal the patterns that occur, by combination of visualizing the data in an informative way and discovering subgroups either in variables or observations that show particular relationships (Alpaydin, 2020, James et al., 2013).

Two widely used sets of unsupervised methods comprise:

- *Dimension reduction methods,* where input variables are combined by a linear or non-linear combination of the original variables. Typical examples are *principal component analysis* (PCA) and *t-Distributed Stochastic Neighbor Embedding* (t-SNE) respectively. These methods form new auxiliary variables in a way that global or local structure in the data is captured by few of these "factors", and thus, visualization or analysis of these new variables enhance the understanding about the underlying structure of the data; and
- *Clustering methods,* where observations are grouped into subgroups in a way that they are as similar as possible to each other within a subgroup and as different as possible from the observations belonging to another subgroup, given a similarity or dissimilarity metric. An example of an unsupervised learning task both in mineral exploration or resource fields would be identifying clusters of drill hole samples with multi-element geochemical data that share similar multivariate features and form distinct, spatially contiguous volumes, also called *domains* in the geostatistical literature, to be further used in resource estimation or geological modeling.

Outputs of an unsupervised learning exercise could be knowledge gain through identified patterns and structures in the data. These patterns could also inform subsequent supervised learning applications. For example, clusters identified during the application of an unsupervised method can be used as labels, or input variables can be analyzed to perform *feature extraction* and *feature selection*. Feature extraction means combining the original variables to form new auxiliary variables, while feature selection refers to discarding variables that are deemed irrelevant to the problem on hand.

The following section provides a list of useful unsupervised and supervised learning tools that can be part of a workflow to exercise ML in mineral exploration and resource field. Specific details of some of these tools are also provided for some of the most relevant methods.

2.4. Machine learning in mineral resource sector

Some MLA can be considered as highly interpretable, e.g. decision trees (DT), but might not be the most accurate approach for most of the problems. Other algorithms such as artificial neural networks (ANN), can be thought as *black box* because of their low interpretability; however, they can capture very complex relationships between input and output data and produce highly accurate predictive results. Based on the availability of the data, nature of the problem and desired outcome, the practitioner must decide which machine learning approach is most suitable for the problem at hand. The following table (Table 2.1) provides a brief summary of the MLA used in mineral resource sector to facilitate the selection of a suitable method. Subsequently, a generic workflow is provided, along with a review of several case studies, to present an overview of the state-of-the-art applications.

Table 2.1 An overview of MLAs that can be used in mineral resource sector for different learning tasks

Method	Common Modeling Approach	Common use in mineral resource sector	Key References
Principal Component Analysis (PCA)	Unsupervised, linearly combines variables to maximize captured variance in principal components, preserve global structure	• Exploratory data analysis • Process discovery in geochemical or geophysical data • Feature extraction • Dimension reduction • Noise reduction in geophysical data • Decorrelate variables for geostatistical simulations.	Theoretical background - Freidman et al., 2013 Application – Grunsky, 2010
Factor Analysis (FA)	Unsupervised, linearly combines variables to maximize common variance in factors, preserve global structure	• Exploratory data analysis • Process discovery in geochemical data • Feature extraction • Dimension reduction	Theoretical background - Freidman et al., 2013

Method	Common Modeling Approach	Common use in mineral resource sector	Key References
			Application – Reimann et al., 2002
Multidimensional Scaling (MDS)	Unsupervised, projects the data into low dimensional space and preserves the pair-wise distances, preserve global structure	• Exploratory data analysis • Dimension reduction	Application (non-mineral resource sector – Shi and Horvath, 2006 (coupled with Unsupervised Random Forest)
Locally Linear Embedding (LLE)	Unsupervised, non-linear projection of the data into low dimensional space, preserve local structure	• Dimension reduction in hyperspectral data	Theory and application (non-mineral resources) - Roweis and Saul, 2000
t-Distributed Stochastic Neighbor Embedding (t-SNE)	Unsupervised, non-linear projection of the data into low dimensional space, preserve local structure	• Exploratory data analysis • Dimension reduction	Theory – Maaten et al., 2008 Application – Case Study I in this document
***K*-Means & -*K*-Medoids Clustering**	Unsupervised, partition data into *K* mutually exclusive groups, minimizes within cluster sum of squares (see text for formula), expected number of clusters should be determined before running the algorithm	• Clustering • Exploratory data analysis • Preprocessing for supervised learning (learn labels and/or automate labeling process) • Geochemical / geophysical mapping • Geological domaining	Application (non-mineral resource sector – Shi and Horvath, 2006 (coupled with Unsupervised Random Forest)
Spectral Clustering	*Unsupervised*, similar to K-Means, but project the data into a lower dimensional space where similarities are enhanced, conducts clustering in the projected space	• Spatial clustering • Geological / geostatistical domaining	Application - Romary et al., 2015 Application - Fouedjio et al., 2017
Hierarchical Clustering	*Unsupervised*, build a dendrogram based on distances between observations, connects closest observations hierarchically, does not require to input *K* before clustering	• Clustering • Exploratory data analysis • Preprocessing for supervised learning (learn labels and/or automate labeling process) • Geological / geostatistical domaining	Application - Romary et al., 2015, 2012
Linear, Multiple, Non-linear Regression	*Supervised*; optimizes weights to minimize, approximates *Y* given *X* as linear (or non-linear in polynomial) combination of *X*	• Estimation • Multivariate estimation for continuous variables in geochemical or remote sensing data	Theoretical background - Freidman et al., 2013
Logistic Regression	Supervised; classification, estimates probability of an observation belonging to a class given *X*	• Prediction for categorical variables • Mineral prospectivity mapping • Predictive lithology mapping	Theoretical background - Freidman et al., 2013 Application – Agterberg et al., 1993

Method	Common Modeling Approach	Common use in mineral resource sector	Key References
K-Nearest Neighbors	Supervised; regression and classification, estimates (in regression) or predicts (in classification) based on *K* nearest (given distance metric) neighbor value (averages and counts votes, respectively)	• Estimation or prediction • Downscaling or upscaling remotely sensed data	Theoretical background - Freidman et al., 2013
Naïve Bayes Classification	Supervised; classification, estimates conditional probability of a class given *X*	• Classification • Mineral prospectivity mapping • Predictive geological mapping	Application – Cracknell and Reading, 2014
Decision Tree and Random Forest	Supervised and unsupervised; recursively split the database by sets of IF - THEN rules to achieve largest decrease in impurity (see text for details), Random Forest is ensemble of Decision Trees	• Classification • Regression • Clustering • Mineral prospectivity mapping • Predictive geological mapping	Theoretical background – Breiman, 2001 Application – Cracknell and Reading, 2013, Case Study II in this document
Artificial Neural Networks (ANN)	Supervised; comprises set of nodes that linearly combines the previous inputs with set of weights and applies a non-linear function to pass a value to the next node. Weights are optimized to minimize a given loss function	• Classification • Regression • Mineral prospectivity mapping • Predictive geological mapping • Pattern recognition, e.g. fault and lineament recognition	Theoretical background and application - Granek, 2016
Support Vector Machines (SVM)	Supervised; finds the hyperplane that provides the largest margin to separate two classes by allowing some degree of error, hence also called soft margin classifier	• Classification • Regression • Mineral prospectivity mapping • Predictive geological mapping	Theoretical background and application - Granek, 2016

In the following section, tools that are deemed to be useful for a general ML workflow in mineral exploration are elaborated in a logical order that is starting from the exploratory data analysis stage, where geologist can formulate relevant questions, identify interesting groupings in the observations and the variables for which unsupervised learning method are useful. This usually is followed by a predictive stage where supervised learning methods are applied to generate a predictive model which is then by followed by evaluation of the results of this model.

Exploratory Data Analysis, Dimension Reduction and Unsupervised Learning

Unsupervised learning methods have close affinity to exploratory data analysis (EDA) as both aim to discover the structures in the data. As such there are lots of tools in the intersection of these two domains. *Dimension reduction* methods can act as exploratory data analysis tools and generally comprise the first step of a multivariate data analysis workflow as they provide the means of projecting the data in a lower dimensional space while preserving much of the information. There are many dimension reduction techniques while some transformed the data linearly e.g. PCA and MDS, some apply non-linear transformations such as LLE, t-SNE (see Table 2.1 and references therein).

Principal component analysis (PCA) is one of the unsupervised learning techniques used for both dimension reduction, visualization and knowledge discovery. It reduces the dimension of the data by creating auxiliary variables, called principal components (PC), which are linear combinations of the original variables. In a N x p data set, where there are N records and p variables, the first principal component of the observation $x_i = \left(x_{i1}, \dots, x_{ip}\right)^T$ can be defined as the linear combination:

$$z_{i1} = \mathrm{w}_{11}x_{i1} + \mathrm{w}_{21}x_{i2} + \dots + \mathrm{w}_{\mathrm{p1}}x_{ip}$$

that maximizes the represented variance, and subject to $\sum_{j=1}^{p} w_{j1}^2 = 1$. The maximization problem is solved by eigen-decomposition of the covariance matrix of the original variables (James et al., 2013). It allows us to visualize the p dimensional data in fewer dimension as most of the variability of the data is captured by the first few principal components. In other words, the first principal components allow us to see what combination of the original variables spread the data most and reveals associations of observations as well as variables linked to each one of those components.

In geoscience data, ideally, each of these new variables can be used to explain different underlying geological processes such as alteration/mineralization, weathering, or metal associations. Therefore, besides its use as a dimension reduction technique, PCA is widely used for knowledge discovery in mineral exploration (Grunsky, 2010).

t-SNE, developed by Maaten et al. (2008), is another way to visualize a high dimensional dataset in a lower dimension by means of a non-linear projection. It calculates the similarity matrix in the form of pair-wise conditional probabilities that represent the likelihood of two points to be neighbors under a t-student distribution centered at one point in the high dimensional space, and then reproduces this structure in the lower dimension. While PCA preserves the global structure of the data, t-SNE aims to reproduce local structures in the original feature space, i.e. t-SNE might reveal different associations. Therefore, it is also useful for dimensionality reduction and knowledge discovery.

There are many other methods for linear and non-linear dimensionality reduction, which can be found in the machine learning literature and seek the same goals as those described for PCA and t-SNE. In summary, these sets of tools help practitioners visualize the data in a simpler, and more informative way and facilitate the detection of important aspects both in observations and variables, and thus, help to *select* or *extract* the best set of features to conduct subsequent studies.

Clustering algorithms comprise another set of unsupervised exploratory tools in a machine learning workflow. They aim to partition the data into groups so that similarity within groups is maximized while dissimilarity between different groups is maximized (James et al., 2013). Similarity or dissimilarity measures depend on the domain of the problem. They are grouped under *unsupervised* methods only because there is no response variable to supervise the learning process,

however, decisions made by the practitioner are crucial to achieve meaningful results and make sense of them. For example, *K-means* is one of the most widely used clustering algorithms that tries to partition the data into *K* clusters that minimize the within cluster differences and mostly uses within cluster sum of squares (WCSS) defined as:

$$WCSS = \sum_{k=1}^{K} \sum_{i \epsilon N} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

where *K* is the total number of clusters as specified by the practitioner before running the algorithm. *K-medians* (Kaufman and Rousseeuw, 1987) is a variant of K-means and it uses the median sample instead of using the centroid of each group to calculate within cluster differences. *Hierarchical clustering* is different in that it does not require to specify the number of desired output clusters prior to running the algorithm. However, it has a tree-like hierarchical representation, called dendrogram, of the resultant groups and the practitioner needs to make a decision about the number of clusters by analyzing the resulting groups. At the bottom of the dendrogram, most similar observations are fused given a similarity metric. These fused observations form mini clusters which are then fused with the most similar mini clusters to form a larger cluster, in something similar to a tree structure. Therefore, towards to the upper parts of the dendrogram, within cluster dissimilarities increase. The practitioner can visualize the dendrogram and decide where to prune to achieve the desired number of *K* clusters (James et al., 2013).

Clustering, like dimension reduction, helps practitioners recognize interesting patterns in the data by grouping them together. These groups then could be used for different purposes, e.g. to formulate research questions and focus the effort in a specific direction or to provide labels to facilitate the subsequent supervised learning phase.

Predictive Modeling and Supervised Learning

Supervised learning methods are divided based on the type of target variable. It is called regression when the target variable is continuous and called classification when the target variable is categorical (James et al., 2013).

Regression methods aims to estimate Y given X by optimizing parameters, β_0 and $\beta_1 \; for \; Y = \beta_0 + \beta_1 X$ such that it minimizes the *residual sum of squares* (RSS) on observations. RSS is defined as:

$$RSS = (y_1 - \beta_0 - \beta_1 x_1)^2 + \ldots + (y_n - \beta_0 - \beta_1 x_n)^2$$

Simple regression can be extended to multivariate problems and is named as *multiple regression;* the generalized formula becomes as follows:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

and aims to minimize

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}\right)^2$$

It can be seen that both simple and multiple regression have linear forms, however, it can be extended to non-linear problems simply by transforming the input *X*, *e.g.* $X^2, \sqrt{X}, logX$.

Although there are many more algorithms for prediction of a continuous attribute, the rest of this section will focus on classification methods as most of the problems in mineral exploration and resource field can be defined as classification problems. For instance, domaining the geological units to facilitate resource estimation, classifying rocks based on different multivariate geoscience

data to improve geological mapping, or deciding if a location is prospective enough to conduct exploration studies for a specific type of commodity.

Artificial neural networks (ANN) are non-linear predictive models formed by layers of nodes connected to each other by sets of weights. Each node is fed by the previous output, that is a linear combination of the values of the nodes on the previous layer (or input data if it is the first layer), and a bias term, which are then passed through a non-linear transformation (Bergen et al., 2019). The aim of the algorithm is to find the set of optimum weights and biases to correctly predict the target variable given input features. ANNs are generally demanding in terms of training data to find the optimum weights and biases which limits their uses in some applications in mineral resource field, e.g. mineral prospectivity modeling where the *true* samples are scarce. ANN can be used for regression or for classification.

Support vector machines (SVM) are soft margin classifiers, which objective is to define the widest soft margins between classes by allowing to make some errors in classification of the observations within the margin. SVM maximizes the margin, M, between classes by allowing the user to tune the tolerance for making errors with a cost parameter C (James et al., 2013). They are originally designed to define linear boundaries (Cortes and Vapnik, 1995), however, non-linear boundaries can be inferred by projecting the original data into higher dimensional space and drawing linear boundaries in this space, which translated into non-linear boundaries in the original features' space. Projecting the data to a higher space can be unmanageable when the data becomes very large and/or the projected space gets very high in dimension. However, SVM simplifies the computation by using inner products of the observations rather than projecting them into the higher dimensional space explicitly. The inner product is further simplified by approximating the distance

between the observations with a kernel (James et al., 2013). This technique is widely adopted for classification studies in mineral resource and exploration.

Random Forest (RF) is an ensemble of many *decision trees* (DT). A DT aims to split the database to achieve maximum purity in each side of the split by searching the best threshold of the best variable. Breiman et al. (1984) postulated that DT starts by splitting the parent node into binary pieces, where the child nodes are purer than the parent node. Optimum splitting criteria are chosen to maximize purity given a purity metric such as Gini index or entropy. Because of their nature, DTs are prone to overfit the data very easily which eventually causes low prediction accuracy for unknown samples (Figure 2.3). One way to avoid this problem is to prune the tree to some extent but this may cause underfitting which results in very coarse classifications (Figure 2.4). The optimum degree of pruning can be found by sensitivity analysis. Random Forest prevents overfitting by randomizing the tree building process. It introduces the randomness to the process in two ways; given that the number of samples in a training set is *N*, *N* number of samples is sampled with replacement, also called bootstrapping in statistics language. This creates a distinct training set for each decision tree in the forest. Secondly, in each split, the algorithm is forced to choose $p_{sub} << p$ out of p variables randomly to split the data so that decisions in the nodes will be partially different in each decision tree. Trees grow without any pruning. The result is the combination of imperfect decision trees. The final decision is made by majority vote (Breiman, 2001).

Measuring performance

Regardless of the modeling approach (unsupervised or supervised), results should be assessed and validated by an expert both qualitatively and quantitatively where possible. The validation or

measure of performance of the learning tasks differs depending whether a problem is cast as unsupervised or supervised. When there is no access to the ground truth, as in unsupervised problems, internal performance metrics can be used as well as qualitative assessment to interpret if results make sense. An example of an internal performance metric is the silhouette coefficient that summarizes how each member of a cluster is close to each other and far from the samples in other clusters (see Palacio-Nino et al., 2019 for a comprehensive review of the methods).

For supervised learning problems where *actual class* information is available, metrics are mostly derived from a confusion matrix which is a table that compares the predicted results against actual results (Table 2.2).

Table 2.2 Outline of a confusion matrix

		Predicted Class	
		Negative	Positive
Actual Class	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Accuracy, $\frac{TP+TN}{ALL}$, true positive rate (or recall, sensitivity), $\frac{TP}{TP+FN}$, precision, $\frac{TP}{TP+FP}$, false positive rate, $\frac{FP}{FP+TN}$, are some of the metrics that are widely used in machine learning.

Another useful metric that is derived from the confusion matrix and used specifically in some binary classification problems such as mineral prospectivity mapping, is called Receiver Operator Characteristics (ROC) (Figure 2.5). Most of the classification algorithms provide class probabilities together with the predicted class labels. These probabilities can be used to determine final classes by thresholding different probability values. For example, when recognizing all the actual true values are critical, a low probability threshold can be chosen to classify samples as

positives. Similarly, when avoiding false positives is important, a high probability threshold can be selected. The ROC curve is derived by plotting the false positive rate on the x-axis and true positive rate in the y-axis for every possible threshold in the probability distribution of the predictions. This plot helps finding an appropriate threshold for the purpose of study. For instance, if the slope of the ROC curve is small (close to horizontal), this indicates very little gain in TPR while increasing FPR, which is not a desired outcome. However, a steep line (close to vertical) indicates a good amount of gain in TPR with respect to increase in FPR. Besides that, the area under this curve, also called AUC, summarizes the total performance of the model and is used to compare different models in binary classification tasks.

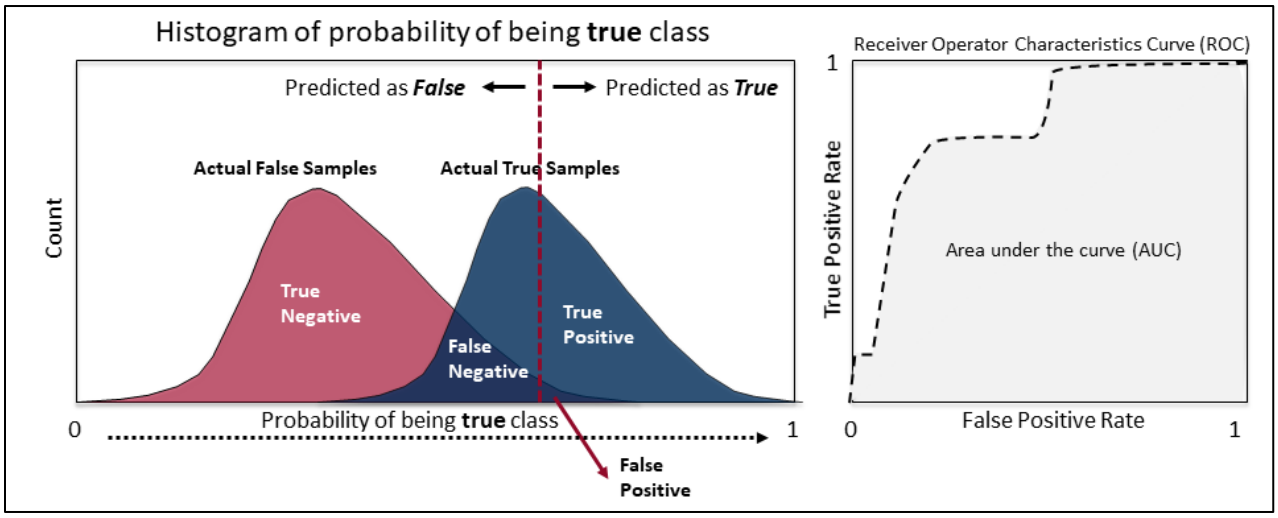


Figure 2.5 A schematic representation of the construction of a ROC curve and AUC metric. Note that, true positive rate (TPR) and false positive rate (FPR) are the function of the specified threshold (red dashed line on the left). Shifting the threshold will reduce the number of false positives and vice versa. Based on the problem on hand, practitioner decides the position of the threshold. ROC curve is generated by plotting TPR vs FPR for every

possible threshold and area under the ROC curve (AUC) is used as a performance metric of a ML classifier. A perfect classifier will AUC value equal to 1, whereas random prediction will have an AUC value of 0.5.

2.5. State of the art application in mineral resource sector

Current research of MLA in the mineral exploration sector is mostly focused on increasing exploration efficiency by increasing high potential search spaces (Hronsky, 2009). Geological maps are essential for this task as they form the basis of any mineral exploration strategy when bedrock exposure is available. However, the new frontiers for exploration include areas under cover for which bedrock geological map is not available in which case sensing tools, e.g. geophysical or geochemical studies, or direct sampling via drilling methods are used to infer the bedrock geology features. Based on the scale of the project, and the conceptual model adopted for exploration, the level of detail required in the map, drilling and geophysical and geochemical survey may change.

Conceptual models adopted for exploration include mineral deposit models, such as, porphyry copper systems (Sillitoe, 2010), epithermal gold deposits (Hedenquist et al., 2000), or mineral system approach (Wyborn, 1994; McCuaig, and Hronsky, 2014). The second most important factor that may impact the mineral exploration efficiency is to utilize all the available geoscience information to evaluate prospectivity of an area. Mineral prospectivity mapping, or mineral potential mapping, approaches aim to develop techniques to facilitate this task.

The following section presents a summary of the state of the art in predictive lithological mapping, mineral prospectivity modeling, and is followed by an overview of the methods that addresses the spatial clustering or domaining needs of the mineral resource sector.

Predictive Lithological Mapping

Geological field studies can be highly time and money demanding, especially for certain parts of the earth, such as Canada's territory north of latitude 60°. Fast and automated, first-pass geological maps, possibly created by MLA, would help focusing the efforts to places where more attention of a geologist is needed, for instance places where geological uncertainty is higher or places with higher mineral potential. This would increase the efficiency of the mapping task in the field.

To assess the potential of using MLA to utilize existing geoscience data to facilitate geological mapping, Cracknell and Reading (2014) compared five MLA, namely, Naïve Bayes (NB), k-Nearest Neighbors (kNN), Random Forest (RF), Support Vector Machines (SVM) and Artificial Neural Networks (ANN), in terms of performance for classification of lithologies by using remotely sensed geophysical data and satellite imagery. They also compared the sensitivities of the methods as a function of spatial distribution of the training data, from highly clustered to dispersed coverage. The training data set, D_{train}, was chosen to cover approximately 10% of the total study area and a 10-fold cross validation on D_{train} was used to tune hyper-parameters of the algorithms. A separate test data set, D_{test}, was used to report accuracy. Their study showed that RF outperformed other MLA with approximately 90% accuracy, while NB has the lowest accuracy (~55%). The performance of the algorithms increased as training data become more dispersed. They also reported that RF has the lowest sensitivity to the hyper-parameter selection.

Hood et al. (2019) showed that RF performs well for predictive lithological mapping with remotely sensed data even in the areas where a transported overburden exists. Their study also showed that selecting (and discarding) the input data and creating auxiliary variables based on expert knowledge can improve the prediction results. In their case, using the PCA transformed data as

input and a reduced set of remote sensing data, resulted in higher accuracies than using all the available data.

Cracknell and Reading (2013) and Kuhn et al. (2019) quantified the prediction uncertainty of RF by using variance and entropy of class probabilities, respectively, and showed that areas with higher uncertainties spatially correlated with geologically complex features such as contact zones or shear zones. Therefore, considering the uncertainty on predictions would help geologist to prioritize field studies to focus on areas where much of the attention is needed. Cracknell and Reading (2013) also introduced a strategy to discard the predictions under certain uncertainty threshold and increase the prediction accuracy.

Harris and Grunsky (2015) presented that a similar approach with combination of lake sediment multi-element geochemistry and geophysical data (airborne total field magnetic and gamma ray spectrometer) could produce meaningful predictive lithologic maps. Validation accuracy for the classification (seven distinct classes) yield approximately 57% accuracy. Although not reported quantitatively, they visually showed that accuracy could be increased by eliminating predictions with lower probabilities than a certain threshold.

Hood et al (2018) used a combination of RF in unsupervised mode and k-means clustering to classify unaltered lithologies into distinct classes by using geochemical data and used these classes as input to link altered versions of these rocks into their equivalent photoliths.

Most of the current applications of MLA on predictive mapping do not take into account spatial autocorrelation explicitly. Some researchers (Kirkwood et al., 2016) suggest that as the explanatory power of the auxiliary variables increase, the importance of spatial autocorrelation

will decrease because the spatial autocorrelation of the variable of interest will be completely captured by these auxiliary variables without explicitly taking them into account.

Talebi et al. (2018) suggested that uncertainties derived from these algorithms cannot be threated as a model of spatial uncertainty. To address this limitation, they combined geostatistical simulations, and MLA. They used turning band simulation algorithm (Emery and Lantuéjoul, 2006) to simulate multivariate geochemical compositions of regolith at unknown locations and utilize RF to produce class probabilities for major crustal blocks of Australian continent for each realization. *Minimum*, *expected*, and *maximum probability* scenarios were presented. They suggested that the resultant *expected probability* model takes in consideration both statistical uncertainties, through bootstrap aggregating mechanism of RF, and spatial uncertainty, through multiple realizations of geostatistical simulations.

Mineral Prospectivity Mapping

Mineral prospectivity modeling is another application of MLA in the mineral exploration sector that attracts much attention among researchers. Mineral prospectivity, or mineral potential and/or mineral favourability, refers to the likelihood that a mineral deposit of interest can be found in a location of interest (Carranza, 2008). Mineral prospectivity mapping aims at ranking a piece of land in terms of mineral prospectivity as a function of available geoscience data that are considered as proxies for the mineral deposit type sought, based on a conceptual model, which are also called evidential features in the mineral prospectivity literature.

Rodriguez-Galiano et al. (2015) compared ANNs, regression trees (RT), RF, and SVM in mineral prospectivity modeling for an epithermal Au district in Rodalquilar, Spain, based on the performance on their accuracy on identifying known deposits, sensitivity to choice of hyper-

parameters, sensitivity to number of training data and interpretability of the models. They combined PCA transformed geochemical data (interpolated with kriging) as a proxy to hydrothermal activity, proximity to certain structures as a proxy to hydrothermal fluid preferential zones, gravity and magnetic geophysical data (interpolated with kriging) as a proxy to certain type of rocks that are considered to be heat source, and hyperspectral remote sensing data as a proxy to hydrothermal alteration types as evidential features. A total of 46 gold occurrences and 57 non-occurrences were selected by stratified random sampling over the study area as target feature, and the above-mentioned MLA was applied to create a final prospectivity map. The 10-fold cross validation was used to choose best parameters for individual models and the models were compared with ROC scores using training points as validation reference. They reported that RF outperformed other methods based on ROC scores whereas the sensitivity of the performance based on hyper-parameter selection was the most robust for RF. Performance of all models decreases gradually with decreasing size of the training data, whereas RF remained the best performer. Carranza and Laborte (2015) also reported that RF is able to produce reliable prospectivity results with as low as 12 known occurrences while most of the other MLA, e.g. ANN, SVM, demands more training data (>20).

A recent study (Sun et al., 2020) compared a set of MLA, namely ANN, RF, SVM, and a deep learning convolutional neural network (CNN), which takes into account spatial patterns in contrast to MLA counterparts, in an area with a total of 118 known Tungsten (W) occurrences as training data (target variable) and 8 evidential layers. These evidential layers included a proximity map to a certain intrusion, density of faults and fault intersections, gravity and magnetic anomaly maps and interpolated W, Fe and Mn geochemical anomaly maps. 10-fold cross validation was utilized to choose the optimum model parameters based on the mean square error (MSE). Based on

ROC/AUC score, RF outperformed the other MLA with an AUC score of 0.96, followed by SVM and CNN with AUC scores of 0.96 and 0.95, respectively. Analysis of the feature importance revealed that previously overlooked Mn occurrences are good predictors for W deposits.

Yeomans et al. (2020), presented a workflow where expert knowledge was imposed to the data-driven model through non-linear transformation of the evidential features using fuzzy set theory functions, for which practitioner adjusts a set of parameters to transform the input data. The subjective nature of the approach is controlled by quantifying the predictive power of the transformed data through ROC/AUC scores. Transformation parameters are adjusted in a way that AUC will be maximized for individual layers. They compared this model with another model fed with the original data without any transformation. Both yielded similar AUC scores (0.96), however, those models differed slightly in terms of spatial distribution of the favorable areas. The model that had applied the fuzzy transformed data yielded new exploration potential areas, i.e. areas with high prospectivity with no known deposits, whereas the model with original data remained relatively limited to areas of known deposits.

Applications in Mineral Resource

In mineral resource field, it is a common practice to identify spatially coherent, statistically similar volumes that are also geologically distinct from other volumes around them. These are called estimation domains in the geostatistical literature (Rossi and Deutsch, 2013), as they improve the performance of estimation techniques. The approach is similar to clustering, where the aim is to define natural groupings in the data. Domaining is usually done manually by an expert, considering few variables, and although it may be supported by geostatistical tools, it is a subjective process.

In this context, Romary et al. (2015) proposed two algorithms: geostatistical hierarchical clustering (GHC) and geostatistical spectral clustering (GSC), where spatial continuity is ensured during the clustering by a proximity condition. Two samples are clustered together only if they are connected on an undirected graph organizing the data. Hierarchical clustering allows the user to inspect the clusters visually in different hierarchical levels to decide the number of desired final clusters, for the case in which this decision must be made for spectral clustering in the beginning and results are sensitive to it. They demonstrated that GHC outperformed other algorithms whereas GSC performed poorly. They noted that one of the limitations of GHC is that it cannot manage large datasets, which can be alleviated with subsampling and using supervised classification afterwards.

Fouedjio et al (2018) demonstrated a method where joint spatial continuity structure, captured by a non-parametric kernel function, is used to create a similarity matrix between samples, which is then used in classical spectral clustering algorithm. They demonstrated the application in an exploration drill hole database of an iron project in Australia with total of 11 geochemical (grades) and 3 mineralogical (mineral abundance and composition) variables. The result is spatially continuous domains with distinct descriptive statistics and experimental variogram results for the variables. As it is the case for most of the unsupervised problems, the decisions are subjective and requires domain knowledge.

A generic workflow for ML applications in mineral resource sector

The active research in mineral exploration and resource sector is focused on three main subjects, predictive lithology mapping, mineral prospectivity modeling and spatial clustering. Predictive classification approaches, such as lithology and prospectivity mapping studies, commonly do not consider the spatial dependency that usually occurs in geoscience data. This means that input

variables are related to the desired output on an individual basis and local, or neighborhood statistics are ignored. Addressing this problem could improve the predictive performance of these methods. It is also important to note that spatial auto correlation must be considered during the validation of the models, otherwise predictive performance metrics are highly overestimated. A minimum generic workflow for a good practice ML application and a brief explanation for each step are presented below (Figure 2.6).

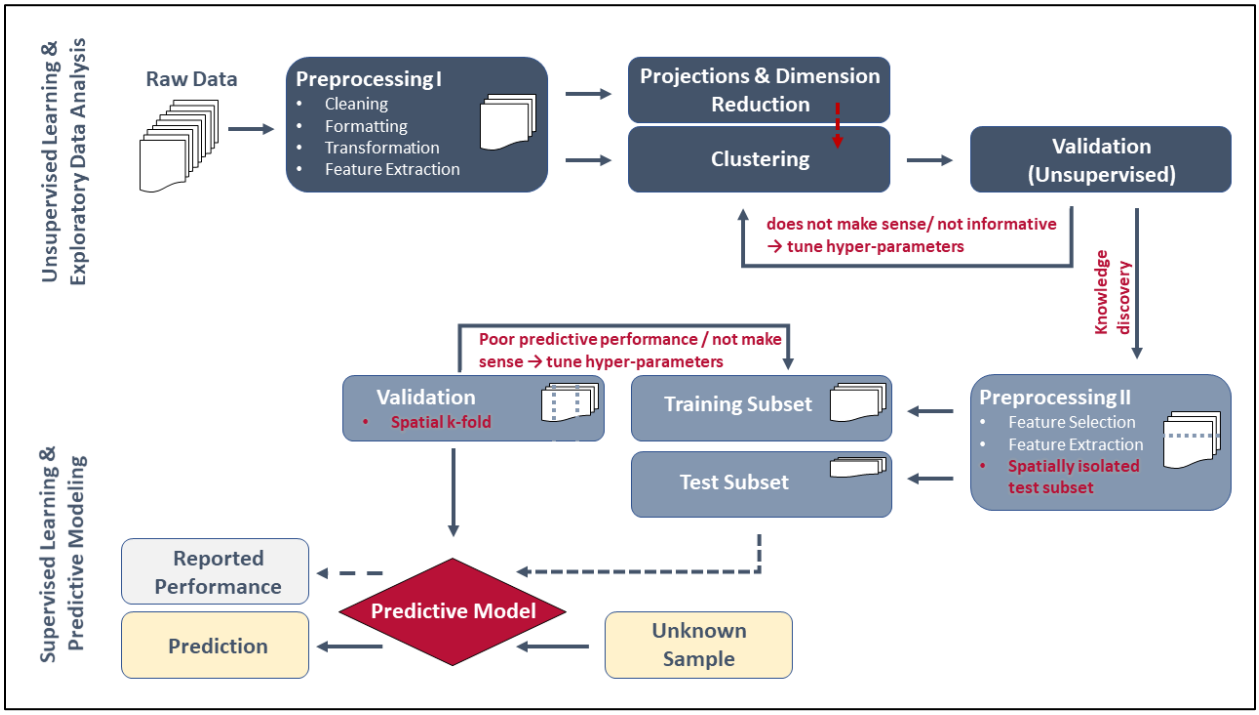


Figure 2.6 A generic workflow for machine learning applications in spatial data. Note the differences in splitting the data into test and validations sets compared to a conventional ML workflow in other fields where spatial autocorrelation is not an issue.

Preprocessing I: This includes (1) detection and removal of erroneous data, (2) converting the data to a format that is suitable to conduct statistical analysis and machine learning applications, that is, usually each observation is located in a row and each variable forms a columns, also called *tidy data* (Wickham, 2014), (3) applying domain specific transformations, e.g. log-ratio transformations for compositional data. It may also include (4) feature extraction through domain

specific transformations, e.g. reduction to pole for magnetic geophysics data, generating proximity maps for relevant geological features or interpolating geochemical data.

Linear / Non-linear projections & Clustering: These include linear and non-linear dimension reduction and clustering methods to project the data into lower dimensional space and visualize the relationship in the data in an informative way. Insights gained here are used to explain phenomenon of interest or formulate relevant research questions.

Validation (Unsupervised Learning): In the case of clustering, results should be validated: (1) qualitatively by interpreting the results and assessing if they are meaningful or useful for the purpose, and (2) quantitatively by using internal evaluation methods. This provides means to tune hyper parameters to achieve meaningful results.

Preprocessing II: This may include (1) feature selection or extraction of new features, and (2) providing new labels, e.g. cluster labels, in light of the knowledge discovered in the unsupervised learning stage, as well as (3) splitting the data into two **spatially distinct** subset for training and testing.

Training and validation: Training a model includes determining the most suitable ML model for the problem and tuning to achieve optimum hyper-parameters through several iterations, training and validations sets. Ideally, iterations should be done after splitting the training data into two spatially distinct subsets. However, *k*-fold cross validation is usually preferred since there are limited number of data. Compare to most other fields, data should be split into spatially distinct *k*-subsets**,** therefore **data should not be shuffled** before splitting, to avoid overestimation of the performance metrics due to unrecognized overfitted model. Besides quantitative validation, a qualitative validation by a geoscientist should be done to ensure geologically sound predictions.

Test & Prediction: This includes testing the model performance on the test subset for reporting purposes. Unknown data can be classified with the predictive model if test results are satisfactory.

Chapter 3 Case Studies

3.1. Introduction

This chapter demonstrates the use of some of the previously described tools through two original case studies which are prepared as standalone manuscripts for peer-reviewed journals. The first study presents a case where knowledge discovery is the ultimate purpose of the analysis, whereas automation and consistency are the main concern for the second case study.

The first case study is titled as "*A combined multivariate approach analyzing geochemical data for knowledge discovery: The Vazante – Paracatu Zinc District, Minas Gerais, Brazil*", co-authored by the two supervisors of the candidate, Dr. Julian M. Ortiz and Dr. Gema R. Olivo, and submitted to the "*Journal of Geochemical Exploration"*. This paper is under review at the time of this thesis submission. A workflow is presented to explore and demonstrate the effectiveness of a combined approach, utilizing linear and non-linear exploratory data analysis tools for knowledge discovery. Principal component analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE) and unsupervised random forest are combined to test a hypothesis related to potential sources for base metal mineralization in the district. This approach revealed patterns of pre-enrichment in the hypnotised source rocks prior to the main mineralizing event, and depletions interpreted to be coeval with the main mineralizing event, suggesting that these rocks could have served as source for the metals. The spatial distribution of identified patterns could assist in identifying new exploration targets.

The second case study is titled as *"On the use of machine learning for mineral resource classification" and* is the result of work supported by a MITACS Acceleration scholarship in collaboration with SRK Consulting Canada. The paper is co-authored by Dr. Julian M. Ortiz,

supervisor of the candidate, Dr. Oy Leuangthong and Dr. Antoine Caté, from SRK Toronto office, and will be submitted to a peer-reviewed journal. We developed a workflow to automate the resource classification task to increase consistency and time efficiency, by using random forest classification in unsupervised and supervised mode, combined with k-medoids clustering and Support Vector Machine (SVM) for post-processing the resulting mineral classification.

3.2. A combined multivariate approach analyzing geochemical data for knowledge discovery: The Vazante – Paracatu Zinc District, Minas Gerais, Brazil

3.2.1 Abstract

The Vazante Group comprises a Proterozoic sequence of carbonate and siliciclastic rocks located in Minas Gerais, Brazil. It is the host of the Vazante-Paracatu Zinc District, including world-class hypogene zinc silicate deposits, in the southern part, and several Pb – Zn sulfide deposits in the northern part, all hosted in dolomitic rocks. A recent study revealed the occurrence of base metal sulfide mineralization that formed prior to the Brasiliano orogenic event in the siliciclastic rocks (Serra do Garrote Formation) that underlie these dolomite-hosted silicates and sulfide deposits. These siliciclastic rocks were considered as potential sources of elements for the hydrothermal fluids that formed the dolomite-hosted deposits, however there was little evidence of depletion of the source rocks during the orogenic event. In this paper, Random Forest is used in unsupervised mode along with t-distributed Stochastic Neighbor Embedding (t-SNE) and principal component analysis (PCA) over a lithogeochemical dataset of samples of the Serra do Garrote siliciclastic rocks collected throughout the basin to provide insights about processes related to the mineralizing system at the Vazante-Paracatu District.

Multivariate analysis reveals that the Serra do Garrote Formation geochemical signature typical of pre-orogenic mineralization. This is characterized by PC1+ with association of Zn, Cd, Cu, Hg, In, V ± Sb, Se, Mo, Re, which corresponds to nearly 60% of the variance in the data. This suggest that these elements are related to a widespread and/or, long-living hydrothermal activity without an efficient mechanism to focus the metal-bearing fluids, causing basin-scale sub-economical Zn and related metal enrichment. Furthermore, PC2 and PC6 distinguish the zones with signatures interpreted to be related to depletion. PC2+ separates Cd-poor sphalerite, which is typical of the generation of sphalerite that occur in zones with textural evidence of remobilisation. PC6+ with association of As, organic carbon and S, identified the zones where sphalerite was leached from the pyrite-rich layers. Areas with multivariate signatures of both pre-orogenic enrichment and syn-orogenic depletion in the source siliciclastic rocks include the Vazante-North Extension and Varginha zinc silicate deposits/occurrence, Ambrósia silicate zinc deposit and the Engenho Velho prospect. Preliminary exploration in this prospect revealed hydrothermal alteration typical of the zinc silicate deposits. This study is pioneering in applying these numerical models in possible source rocks to assist in identifying targets for exploration in basins.

3.2.2 Introduction

Sedimentary basins are important sources for lead and zinc resources globally (Leach et al. 2005). Over the last few decades, our understanding of the mineralizing fluid compositions, hydrothermal alteration, and the processes related to deposition of metals and the formation of these deposits has improved (Large et al. 2005; Leach et al. 2005; Goodfellow 2007). However, little is known about the source rocks of metals. The Mesoproterozoic Upper Vazante Sequence, which comprises intercalated carbonate and siliciclastic rocks located in Minas Gerais, Brazil, hosts the world-class hypogene zinc silicate deposits in the southern sector, and several Pb – Zn sulfide deposits in the

northern sector, accounting for most of the zinc production in Brazil (Olivo et al., 2018 and references therein) (Figure 3.1). The zinc silicate and sulfide deposits formed during the Brasiliano orogenic event and are hosted mainly in the dolomitic rocks of the Serra do Poço Verde and Morro do Calcário formations, respectively. Various researchers have investigated the characteristics of these deposits, the processes by which they may have formed (Monteiro et al., 1999; 2006, 2007; Slezak et al., 2014, Carvalho et al., 2017; Cordeiro et al., 2018; Olivo et al., 2018;), and their petrophysical signatures (McGladrey, 2014; McGladrey et al., 2017) to assist in exploration for similar deposit types. Recently, Fernandes et al. (2019a, 2019b) revealed the occurrence of base metal sulfide mineralization that formed prior to the Brasiliano orogenic event in the siliciclastic rocks (Serra do Garrote Formation) that underlie the dolomite units that host the silicate and sulfide deposits. These siliciclastic rocks were considered as potential sources of elements for the hydrothermal fluids that formed the dolomite-hosted deposits (Olivo et al. 2018). However, based mainly on mineralogical studies and preliminary statistical analysis of the lithogeochemical data, Fernandes et al. (2019a, b) have not found any compelling evidence of possible depletion of the siliciclastic rocks throughout the basin to support the interaction of these rocks with the orogenic hydrothermal fluids.

Some of these previous studies generated a coherent lithogeochemical database analyzed using the same methods and comprising multi-elemental compositions of various units of the Upper Vazante Sequence and the zinc deposits (Carvalho et al., 2017; Fernandes et al., 2019a; Fernandes et al. 2019b). This study aims to apply multivariate and advanced statistical methods to the lithogeochemical database on the Serra do Garrote Formation (Fernandes et al., 2019a and 2019b). This is a pioneering approach that has direct implication to metal mobility in basins and therefore to exploration in similar geological settings, as it allowed for identification of widespread metal

enrichment at early stages of basin evolution and zones of depletion in late stages. The results support the hypothesis that pre-orogenic sub-economic mineralization at Serra do Garrote Formation could be a potential source of metals for syn-orogenic Zn (-Pb) deposits hosted in the dolomitic rocks.

3.2.3 The Vazante Upper Sequence: geological setting and base metal mineralization

The Vazante Group that hosts both zinc silicate and sulfide deposits, occurs as an N-S trending, 250 km long belt located in the southern region of the Brasilia Fold Belt, along the western margin of São Francisco Craton (SFC) in Brazil (Figure 3.1). The Vazante Group was divided in Neoproterozoic Lower Vazante and Mesoproterozoic Upper Vazante sequences by Misi et al. (2014) (Figure 3.2) (Olivo et al., 2018; Fernandes et al., 2019b and references therein).

The siliciclastic and dolomitic rocks of the Mesoproterozoic Upper Vazante Sequence were deposited in the western margin of the São Francisco Craton between 1300 and 1000 Ma in a rift environment at shallow water conditions, where the detrital material derived mainly from a Paleoproterozoic arc-type rock (Fernandes et al., 2019a). The onset of the Brasilia Orogeny started with the convergence of the western edge of the SFC with Neoproterozoic Goiás Magmatic Arc between 850 – 750 Ma and resulted in the formation of Brasilia Fold Belt, approximately 1000 km long fold and thrust belt (Dardenne, 2000 as cited in Fernandes et al., 2019a). During the orogeny, the Upper Vazante sequence was thrust above to the Neoproterozoic Lower Vazante Sequence (Misi et al. 2014). Review of the regional and local geology is provided in Dardenne (2000), Monteiro et al. (2006), Misi et al. (2014), Olivo et al. (2018) and Fernandes et al. (2019b).

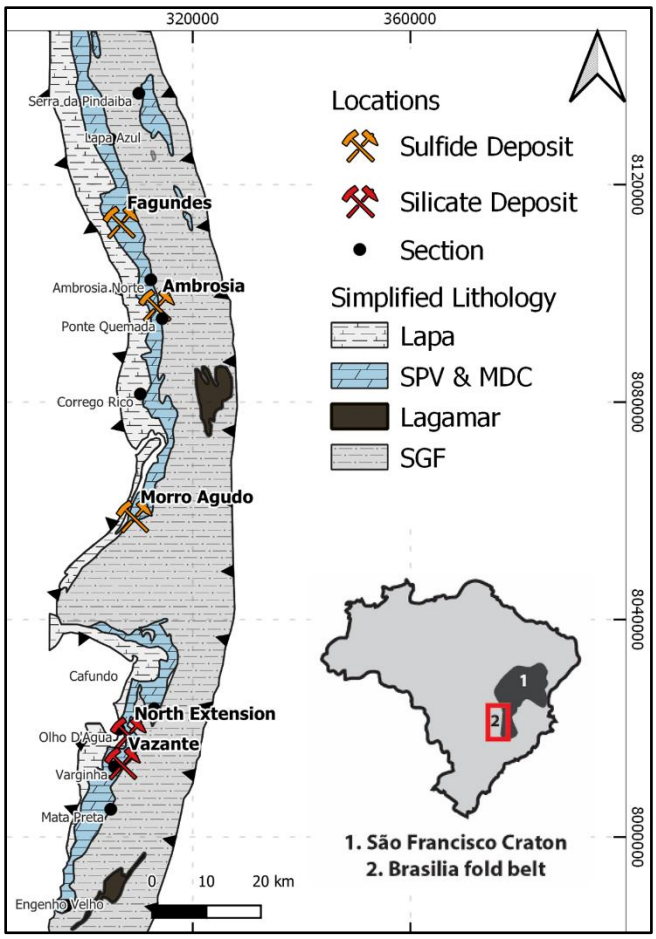


Figure 3.1 Geological map of the Vazante Sequence in the western margin of the Sao Francisco Craton, central Brazil, and location of the known deposits (modified after Fernandes et al., 2019a), and sampling sections (black circles). Sulfide deposits; Morro Agudo, Ambrósia, Fagundes are located in the central to northern part and Vazante and North Extension, silicate deposits are located in the southern part. SPV: Serra do Poço Verde Fm., MDC: Morro do Calcario Fm., SGF: Serra do Garrote Fm.

Sulfide deposits in the belt (Figure 3.1), in which the main ore minerals are sphalerite and galena, include *Morro Agudo Zn – Pb underground mine* (20 Mt resource @5% Zn, 2% Pb), *Ambrósia Zn – Pb open-pit mine* (2.15 Mt resource @5.08% Zn, 0.16% Pb) and *Fagundes deposit* with no published resources (Fernandes et al., 2019a). The main ore mineral in the *Vazante and North Extension underground mines* in the south, on the other hand, is the hypogene willemite, which is a zinc silicate mineral, Zn_2SiO_4. The combined resources are 37.6 Mt @ 19.78% Zn, 0.48% Pb, 29.94 g/t Ag (Fernandes et al., 2019a) (Figure 3.1).

Sulfide mineralization in the north is hosted by the Morro do Calcário Formation (Figure 3.2). It has been suggested that the base metals were transported by hot metalliferous fluids, with salinities between 11 to 32 wt. % eq. NaCl, temperatures ranging from 80 – 300 °C and low contents of reduced sulfur, and the ore minerals precipitated due to mixing of these metalliferous fluids and sulfur-bearing seawater/connate waters during the Neoproterozoic (Misi et al., 2014; Monteiro et al., 2006).

Unlike the sulfide ore, silicate mineralization in the south is hosted by the Serra do Poço Verde Formation (Figure 3.1 and Figure 3.2). It is suggested that ore elements in silicate deposits were transported by fluids with similar composition as the sulfide deposits. In the silicate deposits, the metals precipitated by mixing with an oxidizing, S-poor meteoric water, possibly carrying SiO_2, which have caused abrupt change in temperature, fugacity of oxygen and pH conditions leading to the formation of the willemite (Monteiro et al., 2007; Olivo et al., 2018).

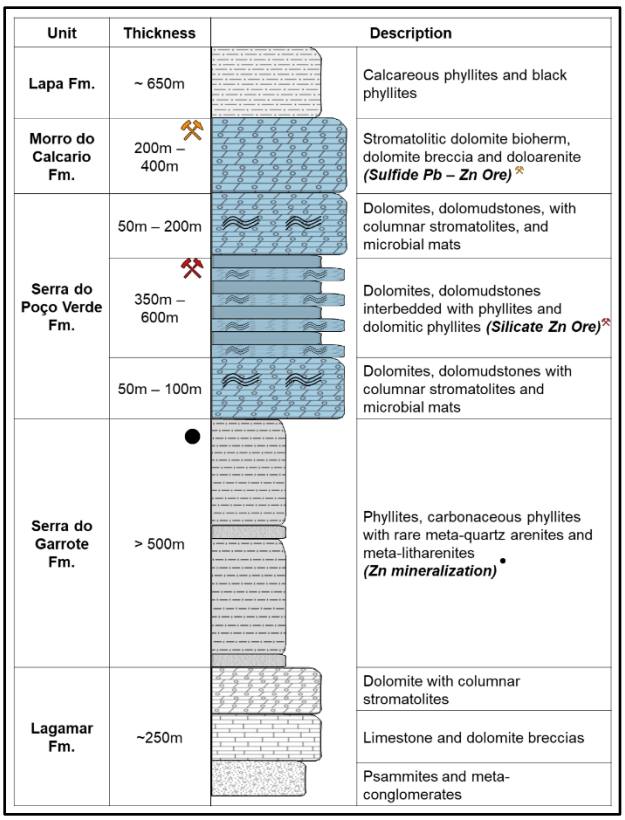


Figure 3.2 Stratigraphic section of the Mesoproterozoic Upper Vazante sequence and the locations of the deposits / mineralization (modified from Dardenne, 2000; Neves, 2011: Misi et al., 2014).

Recently, Fernandes et al. (2019a), reported the occurrence of base metal sulfide mineralization in the siliciclastic rocks of the Serra do Garrote Formation (SGF) (Figure 3.2). The SGF comprises mainly of carbonaceous phyllites and phyllites, grouped in three distinct subunits; SG1, SG2, and SG3 based on molar Al/Ti ratios (Fernandes et al., 2019b). This base metal mineralization is interpreted to have occurred prior to the main orogenic event and is characterized by disseminated sulfide in folded laminations and veins containing pyrite, sphalerite, quartz, and chlorite. This mineralization is hosted mainly in carbonaceous phyllites (> 1% organic carbon) belonging to the SG1 and SG2 subunits, and the SG3 subunit is poorly mineralized. The geochemical signature of the samples with higher Zn values is accompanied by higher concentrations of Cd, Cu, Hg, In, V ± Sb, Se, Mo, Re compared with the background host units. These ore-related elements are similar

to those reported in the carbonate-hosted orebodies (Monteiro et al., 2007, Slezak et al., 2014, Cordeiro et al., 2018). This similarity lead to the suggestion that Serra do Garrote Formation could be a potential source of metals that were subsequently leached by hydrothermal fluids and concentrated economically in the carbonate units during the Brasiliano orogenic event (Olivo et al., 2018) or has a common source with the deposits above (Fernandes et al., 2019a). However, only preliminary statistical analysis of the geochemical data was applied in the previous studies to evaluate the signature of enrichment and possible depletion. Mineralogical evidence of depletion includes the occurrence of Cd-rich sphalerite in zones of transposition along the main foliation and the occurrence of pyrite-rich zones where the sphalerite and galena inclusions were partially leached (Figure 3.6 I-K in Fernandes et al., 2019b).

3.2.4 Methods

Analytical methods and database

A geochemical database with 203 samples from SGF is used in this study (Fernandes et al., 2019a, 2019b). The major oxides were analyzed by X-ray fluorescence (XRF) on fused rock powders. Trace elements were analyzed by a combination of digestion methods, such as acid digestion of fused bead, four-acid digestion and aqua regia digestion and with two different finishing methods, namely inductively coupled plasma mass spectrometry (ICP-MS) and inductively coupled plasma-atomic emission spectrometry (ICP-AES) (See Table 3.1 for breakdown of the elements). A LECO analyzer and furnace measured total carbon, organic carbon and sulfur (Fernandes et al., 2019b). In total, 68 variables are available (Table 3.1).

Table 3.1 Summary of whole rock analytical methods used in this study (Fernandes et al., 2019a, 2019b).

Data	Details
Major Oxides (%) - XRF	Al_2O_3, BaO, CaO, Cr_2O_3, Fe_2O_3, K_2O, MgO, MnO, Na_2O, P_2O_5, SiO_2, SrO , TiO_2
Multi Element (ppm) – Acid digestion of fused bead, ICP-MS finish	Ba, Ce, Cr, Cs, Dy, Er, Eu, Ga, Gd, Ge, Hf, Ho, La, Lu, Nb, Nd, Pr, Rb, Sm, Sn, Sr, Ta, Tb, Th, Tm, U, V, W, Y, Yb, Zr
Multi Element (ppm) – *Aqua Regia, ICP-MS finish*	As, Bi, Hg, In, Re, Sb, Sc, Se, Te, Tl
Multi Element (ppm) - Four-acid digestion, ICP-AES finish	Ag, Cd, Co, Cu, Li, Mo, Ni, Pb, Sc, Zn
LECO Analyzer (%)	Total carbon (C), Total Sulfur (S), Total organic carbon (TOC)

Data Preprocessing

Geochemical data usually contains censored values. A censored value occurs when the concentration of an element for a sample is outside of the detection limits of that element for a particular analytical method. For statistical analysis, the samples that contain censored values in some of their variables may be either completely be removed from the database, which may cause loss of a substantial amount of information, or some alternative imputation methods can be used. In this study, the elements with more than 60 % of their samples censored or missing were removed (Ag and Ge). The remaining 66 variables, that contained missing values, were imputed with a method that is suitable for the compositional data according to Hron et al., 2010. The method uses the k-nearest neighbor approach based on Aitchison distance (Aitchison et al., 2000). A function called 'robCompositions' is available in the R programming language which is used in this study (Hron et al., 2010; Templ et al. 2011). The Aitchison distance is defined as:

$$d_a = (x, y) = \sqrt{\frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2}$$

where d_a is the distance between samples x and y, D is the total dimension of the samples (total variables) and $x_{i,j}$ and $y_{i,j}$ are the ith and jth variables of the sample.

The second preprocessing step is related to the nature of the compositional data, for which the data are strictly positive and carry only relative information, i.e. reported elements represent the part of a whole or proportion, but not the absolute value (Van den Boogaart et al., 2013; Pawlowsky-Glahn et al. 2015). Since variables must sum to a whole (a constant), increasing the value of one of the variables inevitably decreases the others, regardless of they have an inverse relationship or not, which causes spurious correlations between variables (Pearson, 1897; Chayes, 1960). Therefore, ratios are needed to conduct classical statistical analysis. Ratios, on the other hand, (1) are not symmetric, in other words, changing the position of the numerator and denominator does not result in symmetric values around a specific value, and (2) they are strictly positive numbers. In order to address these limitations, using log-ratios was proposed by Aitchison (1982).

This study applied the centered log-ratios (clr) for principal component analysis, which alleviates the above-mentioned limitations and improves the identification of the multivariate associations (Carranza, 2011). The generic formula of the clr transformation is given below:

$$clr(x) = \left[ln\frac{x_1}{g(x)}; \ldots; ln\frac{x_D}{g(x)}\right]$$

where $g(x)$ is the geometric mean of the composition.

Unsupervised Random Forest

Random Forest (RF) is a machine learning method which is an ensemble of many Decision Trees, and widely used in supervised mode to conduct classification tasks. Although less common, it can be used in the unsupervised mode as an exploratory analysis and clustering tool thanks to its capability to produce a proximity matrix (Breiman, 2001). This proximity matrix can then be converted to a distance matrix that allows conducting clustering, or as in our case, feeding other manifold learning methods such as t-SNE for exploratory visualization.

This method was chosen because one of the biggest advantages of using Random Forest in unsupervised mode (URF) is the fact that the distance matrix between samples is able to assimilate categorical and numerical data. This is an important feature for geoscience data, especially for mineral exploration since it allows us to assimilate the qualitative and categorical descriptions of a geologist such as the color of a soil sample, sorting degree of a sandstone package or grain size of an intrusive rock with the quantitative analysis results. The other important advantage of using URF to create a distance matrix is the fact that it does not need to transform the data to a common scale, meaning that it can assimilate variables with different magnitude of scales without needing a pre-scaling or transformation of the data (Breiman, 2001).

URF pair-wise distance matrix between samples was generated by using an open-source function which is available for R programming language (Shi et al., 2006). This distance matrix was then used for t-SNE projection.

Visualization using unsupervised RF and t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE, developed by Maaten et al. (2008), is a way to visualize a high dimensional dataset in a lower dimension. It aims to capture the local structure of the data, whereas principal component analysis aims to capture the global structure. t-SNE calculates the similarity matrix in the form of pair-wise conditional probabilities that represent the likelihood of two points to be neighbors under a t-student distribution centered at one point in the high dimensional space, and then reproduce this structure in the lower dimension. This method was applied to reveal some possible interesting local structures that may be scrutinized in detail with principal component analysis.

The method has implementations in many platforms; the Python implementation was used in this study. It is possible to use the original dataset from which the pair-wise distances are calculated with the desired metric, such as Euclidian, hamming or Mahalanobis among many others, or a precomputed distance matrix can be introduced. In this application, we provided a distance matrix precomputed by the URF for the benefits mentioned in the previous section.

Principal Component Analysis

Principal component analysis (PCA) can be used as an exploratory data analysis tool for analyzing the multivariate geochemical data since it reduces the dimensions of the data by creating new auxiliary variables which are a linear combination of the original variables. In geoscience data, ideally, each of these new variables can be used to explain different underlying geological processes such as alteration/mineralization, weathering, metal associations etc. (Grunsky, 2010). Therefore, PCA can be used for both the purposes of dimension reduction and factorization to understand the underlying processes which separate the different classes.

In this study, PCA was applied to clr transformed data to interpret the processes and visualize the relationship between samples and variables in biplots by using the *prcomp* function of the R statistical programing language.

The flow chart for the various stages of the numerical modelling is shown in Figure 3.3.

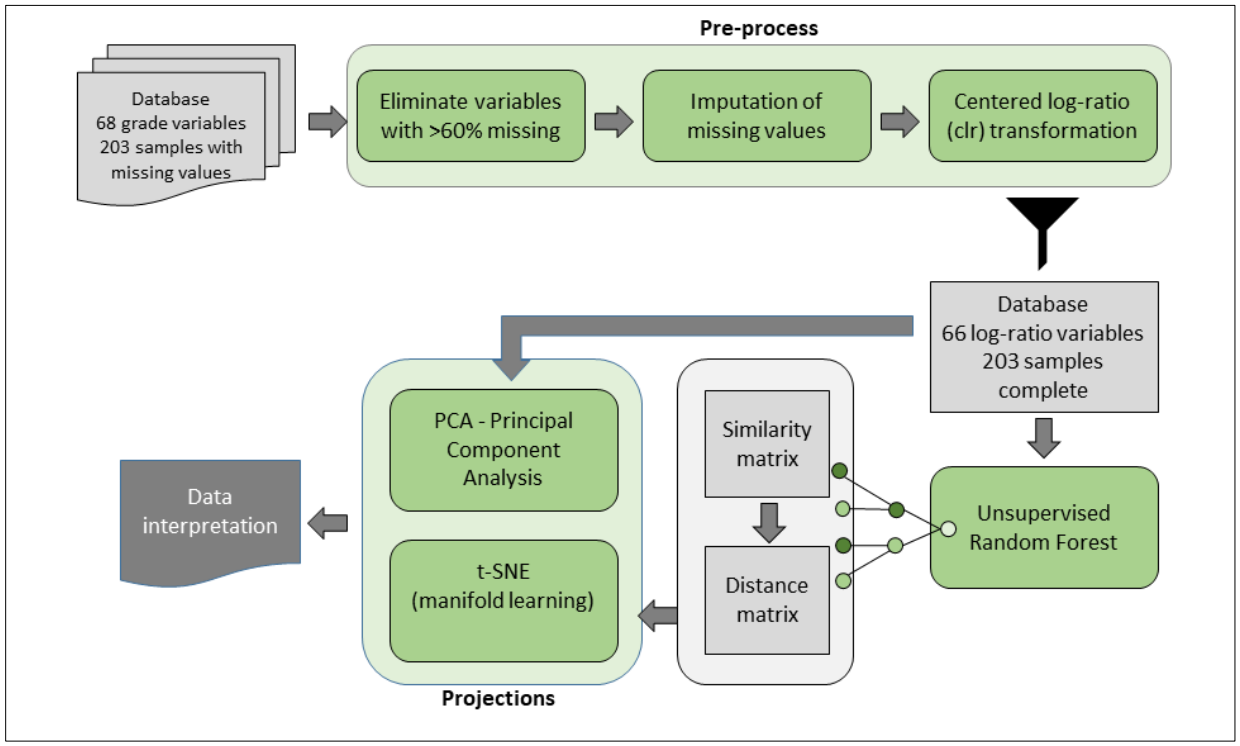


Figure 3.3 Flow chart of steps used in this study. It comprises the following steps: (1) pre-processing the geochemical data; (i) removing the variables that have a high amount of censored or missing values, (ii) imputation of remaining censored or missing values, (iii) centered-log ratio transformation, (2) generating distance matrix with URF, (3) Multivariate analysis with PCA and nonlinear manifold learning, (4) Data interpretation.

3.2.4 Results

The SGF dataset was analyzed with the methods described in the previous section and took in consideration the subunits identified by Fernandes et al., (2019b; SG1, SG2, SG3), which are used as main classes and discriminated in the plots. The relationship among the variations observed in

the geochemical data, lithofacies types and sample locations were evaluated in order to assist in identifying the processes related to this mineral system and improve exploration strategies in this geological setting.

Visualization with t-SNE and unsupervised random forest

The t-SNE projection of all the samples in the database is shown in Figure 3.4A, and Figure 3.4B-L highlights the samples based on the location of the occurrences from north to south (See Figure 3.1 for the locations of the occurrences).

Two distinct main clusters can be seen in Figure 3.4A. Cluster 1 contains almost all of the SG3 samples (blue) with a few of SG1 (red) and SG2 (green) samples, and Cluster 2 comprises SG1 and SG2 samples and only one SG3 sample. SG1 and SG2 samples in Cluster 2 are clustered away from each other (Figure 3.4A).

All the samples from *Lapa Azul* and *Serra da Pindaiba* occurrences (Figure 3.4B-C) are sandstone-siltstone facies from subunit SG1, and clustered within Cluster 1, which is distinct from the majority of the SG1 samples that plotted in Cluster 2.

The samples from *Ambrósia & Ambrósia North, Ponte Quemada, Córrego Rico* (Figure 3.4D, E, F, respectively), which occur in the northern part of the study area, show similar behavior. Almost all SG1 and SG2 samples from these occurrences are in Cluster 2, and SG3 samples are located in Cluster 1. However, the SG1 unit samples of *Ponte Quemada* (Figure 3.4E) are slightly clustered away from the rest of the SG1 samples within the main cluster.

The samples of *Cafundo* (to some extent), *North Extension, Olho D'Agua & Varginha, Mata Preta* and *Engenho Velho* (Figure 3.4G-L) occurrences, which are located in the southern part, behave

similar to each other and different than the other occurrences from the northern part. SG3 samples are clustered in the Cluster 1, similar to all occurrences, whereas SG1 and SG2 samples are split between Clusters 1 and 2. The few samples from Morro Agudo, which belong to SG2 and SG3, plotted in Cluster 2.

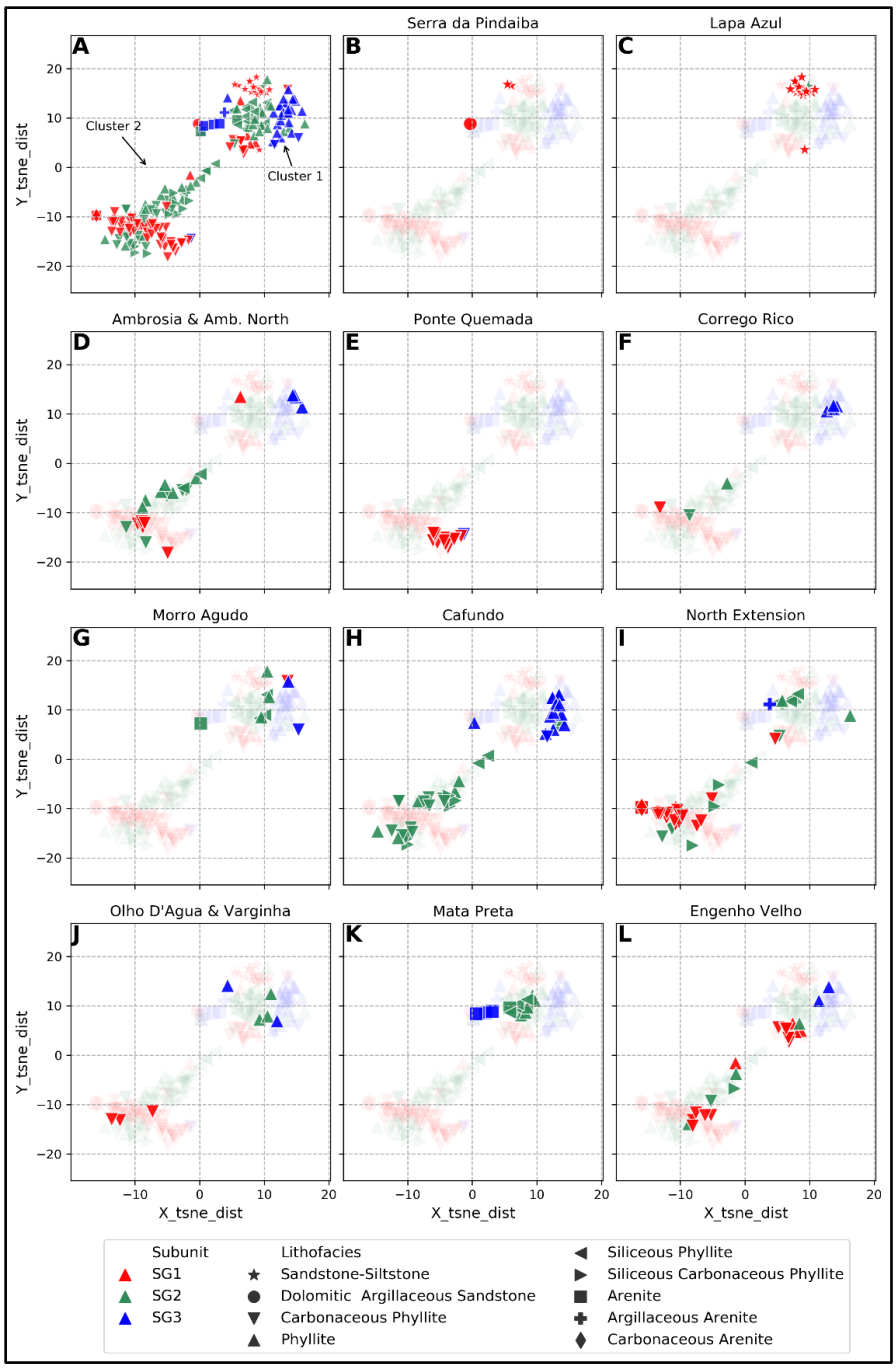


Figure 3.4 t-SNE plots of the samples. 3.4A presents all the samples, whereas 3.4B – L shows the samples from individual occurrence sections from the north (3.4B) to the south (3.4L). Notice that, major clusters (3.4A) corresponds to the separation in the PC1 axis (Figure 3.6). The aim of the t-SNE projection here is to extract the local structure of the data effectively in a 2D plot. Although t-SNE itself cannot answer the reason for a

structure, the underlying processes related to these structures can be explored through PCA. Interesting patterns are mentioned in the text.

Principal Component Analysis

PCA built the correlation matrix between the 66 variables (centered log-ratios of geochemical concentrations), performed an eigen-decomposition and then the principal components were obtained by projecting the samples on the eigen vectors, which are sorted from highest to lowest. The amount of the explained variance by each principal component is shown in Figure 3.5. 80% of the variance is captured by the first six PCs, with 60% of the variance captured by PC1. PCs with low eigen values were not discussed as they do not represent a significant portion of the total variance.

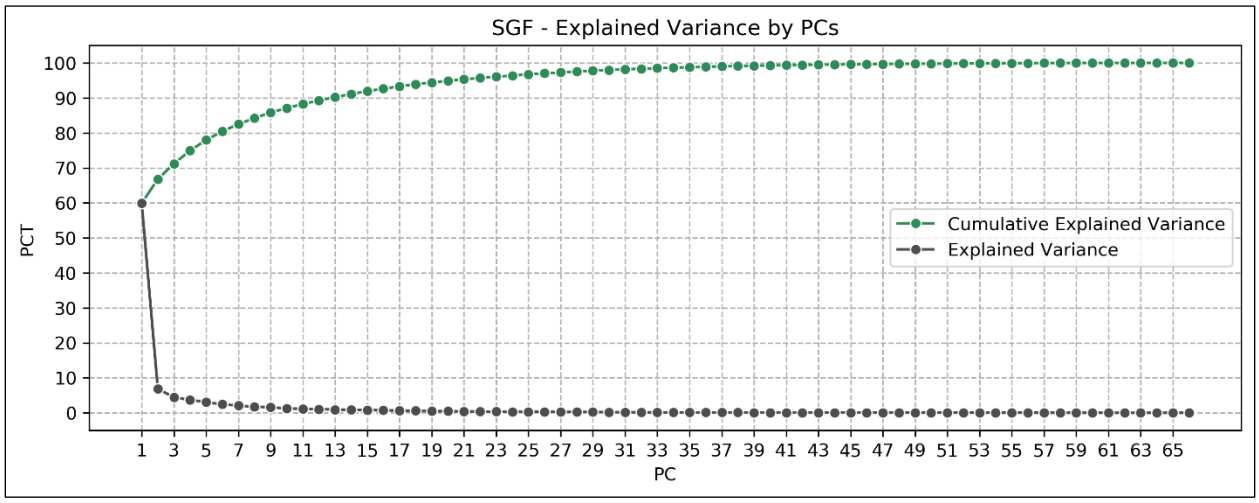


Figure 3.5 Individual and cumulative explained variance (%) for SGF dataset by each principal component. The first 6 PCs captures nearly 80% of the total variation in the dataset.

The element associations of the first 6 principal components are presented in Figure 3.6, PC1,2 and 6 scores of the samples are given in spatial context in Figure 3.7 to Figure 3.9, and PC1 vs PC2 biplot is presented in Figure 3.10. Coincidently, PC1 represents the two major clusters that were captured and clearly visualized by the t-SNE plots (Cluster 1 and Cluster 2 in Figure 3.4A).

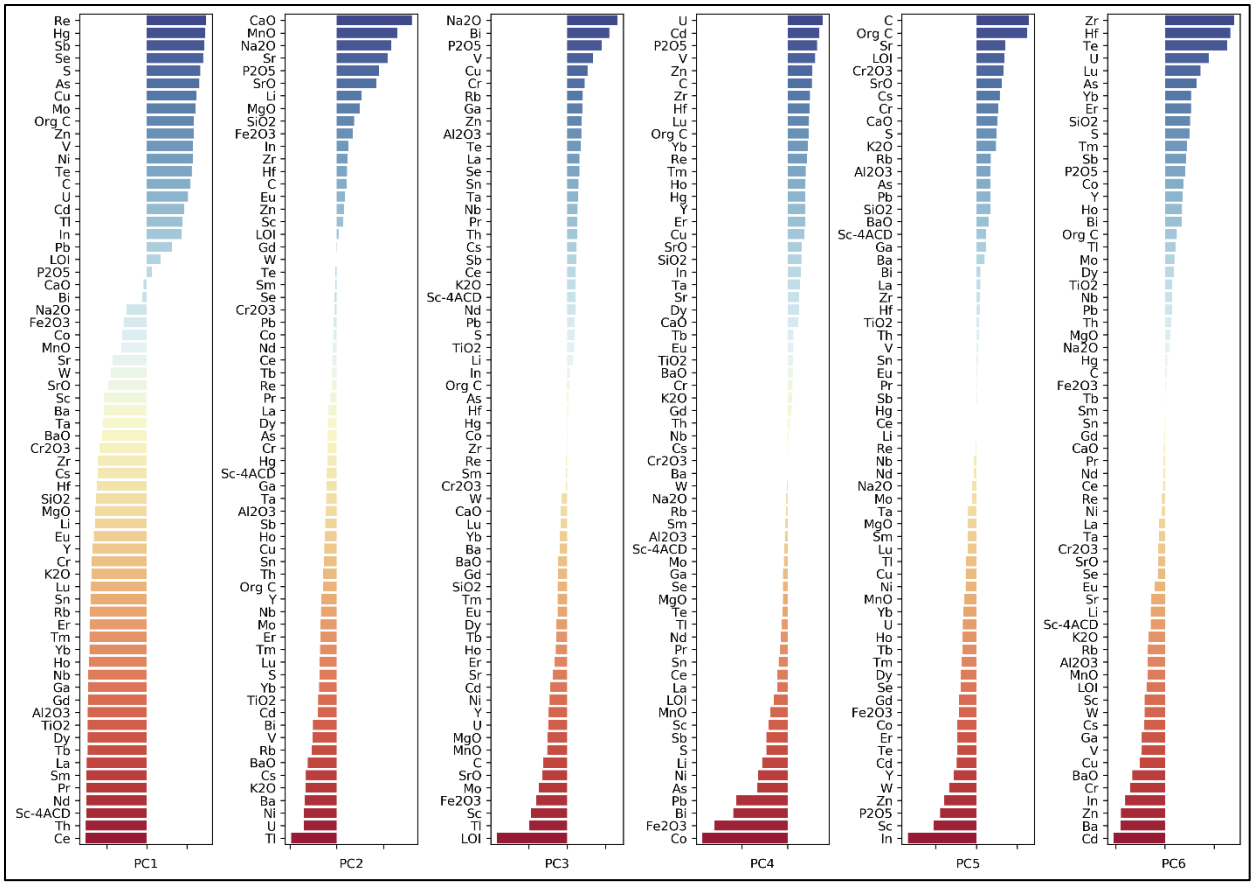


Figure 3.6 Element loadings on the first 6 PC which corresponds to approximately 80% of the total variation.

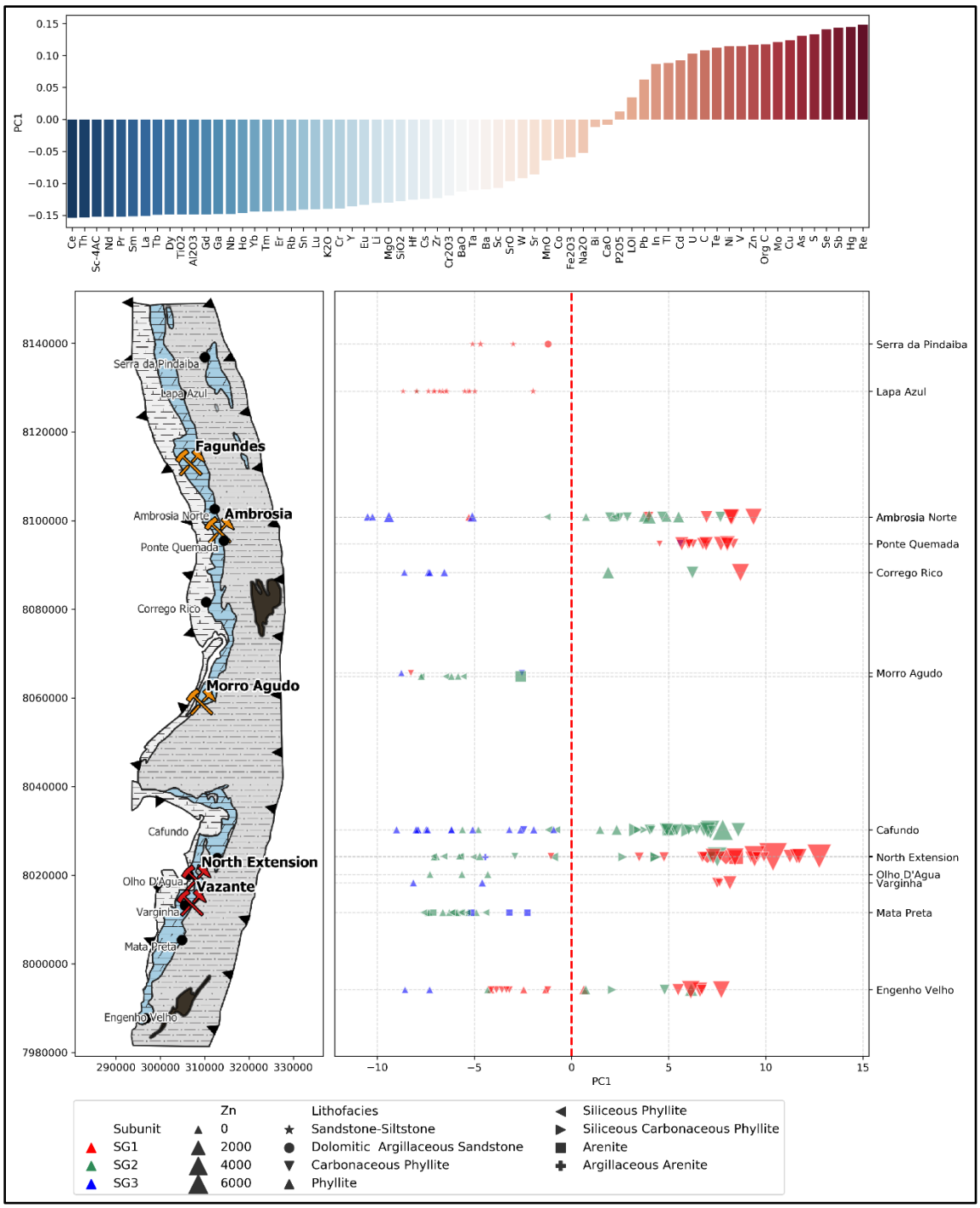


Figure 3.7 Map shows the PC1 distribution of the samples in spatial context together with element loadings in the PC1, and the relative sample zinc contents. The geological map on the left is modified from Fernandes et al., 2019b.

Similar to t-SNE projection, almost all the SG3 samples have negative PC1 (PC1-, with elemental association of REE, Th, Sc, Al_2O_3, TiO_2, Ga, Nb, K_2O, MgO, SiO_2, Hf, Cs, Zr, Cr, Ba), whereas the majority of the SG1 and SG2 samples have positive (PC1+; with Re, Hg, Sb, Se, S, As, Cu, Mo, organic carbon, Zn, V, Ni, Te, C, U, Cd, Tl, In, Pb; Figure 3.7), and all SG1 samples from the *Lapa Azul* and *Serra da Pindaiba* occurrences have PC1-, as opposed to the majority of the SG1 samples (Figure 3.7 and Figure 3.10). The samples from *Ambrósia, Ambrósia North, Ponte Quemada, and Corrego Rico* also show similar behavior in the biplot and t-SNE projection, that is, SG1 and SG2 samples are clustered together, in PC1+, whereas SG3 samples are associated with PC1- (Figure 3.7 and Figure 3.10).

PC2 comprises higher loading on CaO, MnO, Na_2O, Sr, P_2O_5, Li, Mg, SiO_2, Fe_2O_3, In, Zr, Hf, C, Eu, Zn, Sc and LOI in the negative axis and Tl, U, Ni Ba, K_2O, Cs, Rb, V Bi, Cd, TiO_2, Yb, S, HREE, Mo, Nb, Y organic carbon, Th, Cu, Sb, Al_2O_3, Ta, Ga, Sc and Hg in the positive axis (Figure 3.8). *Ponte Quemada* samples have a distinct signature in PC2- direction, with higher loadings in Cd, Ni, V and U (Figure 3.10).

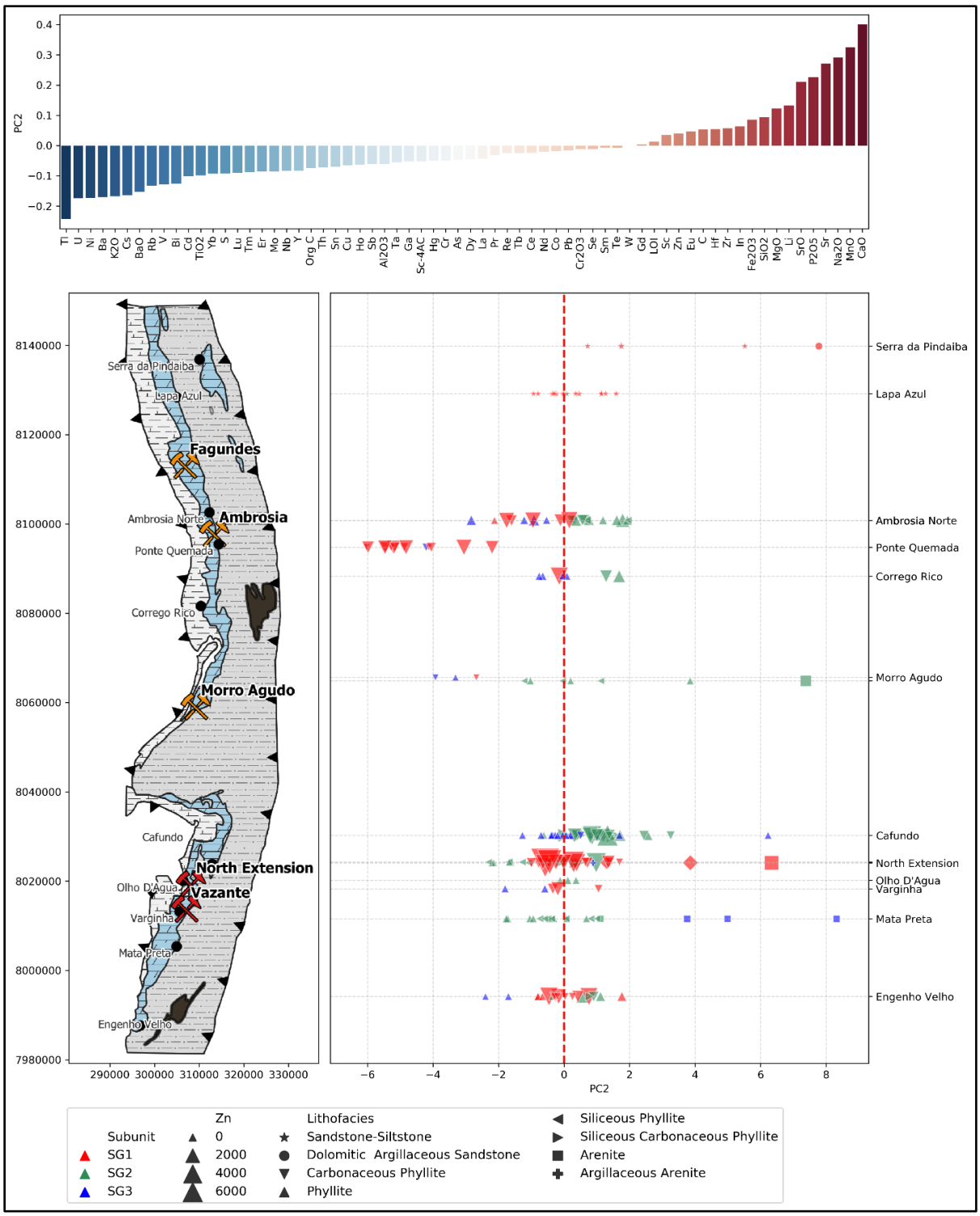


Figure 3.8 Map shows the PC2 distribution of the samples in spatial context together with element loadings in the PC2, and the relative sample zinc contents. The geological map on the left is modified from Fernandes et al., 2019b.

Conformable with the t-SNE plots, the samples from occurrences that are located in the central to southern part, i.e. *Morro Agudo, Cafundo, North Extension, Olho D'Agua & Varginha, Mata Preta* and *Engenho Velho*, show a different behavior than the northern ones. Whereas SG3 samples are still clustered together in PC1-, SG1 and SG2 samples are clustered either in PC1- or PC1+ (Figure 3.7 and Figure 3.10). As for PC2, the samples from these locations are spread between PC2+ and PC2-; the arenites have the highest PC+ values (up to +8), and the phyllites values range between -4 to +4 (Figure 3.10).

PC3 comprises loss on ignition (LOI), Tl, Sc, Fe_2O_3, Mo, SrO, C, MnO, MgO association as dominant loadings in the negative axis and Na_2O, Bi, P_2O_5, V and Cu in the positive axis whereas PC4 comprises, U, Cd, P_2O_5, V, Zn, C, Zr, Hf, Lu and organic carbon in the positive axis and Co, Fe_2O_3, Bi and Pb in the negative axis. PC5+ is dominated by carbon and organic carbon and PC5- is comprised of In Sc, P2O5 and Zn. Their spatial distribution is considered as homogenous and not distinctive between the sections studied in this work (Appendices A, B, C).

Further investigation of the smaller principal components revealed that PC6- (Figure 3.9) which comprises Cd, Ba, Zn, In, Cr, BaO, Cu, V, Ga, Cs, W, Sc, LOI, MnO, Al_2O_3, Rb, and K_2O (from higher to lower loadings), separates most of the SG3 samples from the SG1 and SG2 samples that are located on PC1- field (Figure 3.11). In the PC1+ field, PC6 distinguishes two major groups: PC6- associated with Zn, In and Cd, typical sphalerite composition in the SGF occurrences, as well as V and Cu, and PC6+ with As, organic C, Mo, Pb, S, Sb, Te, Tl, and U. The spatial distribution of PC6 is presented in Figure 3.9.

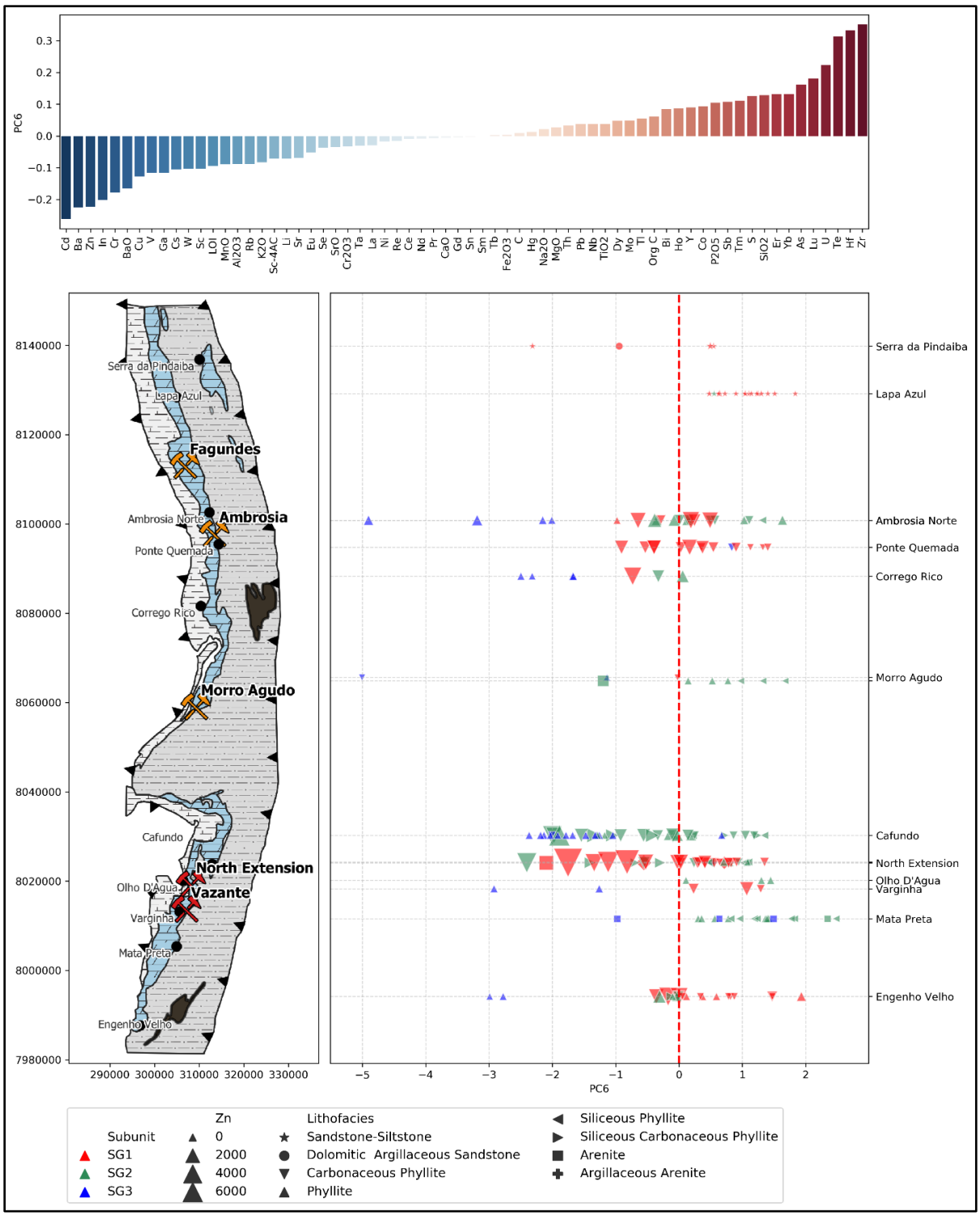


Figure 3.9 Map shows the PC6 distribution of the samples in spatial context together with element loadings in the PC6, and the relative sample zinc contents. The geological map on the left is modified from Fernandes et al., 2019b.

3.2.5 Discussion

Source of Metals: enrichment vs depletion

In regional geochemical surveys, the first principal components usually capture the element associations of lithological variations, whereas the rare processes, such as hydrothermal events (including enrichment and depletion), can be captured by smaller principal components (Grunsky 2010). Fernandes et al.(2019b) used molar Al/Ti ratios to classify subunits based on the assumption that these elements are relatively immobile during metamorphism and alteration. Therefore, distinct ratios of these elements in sedimentary rocks would reflect the difference in their protoliths and/or source rocks. Multivariate analysis conducted in this study supports this subunit classification scheme. Even though major clusters, *Cluster 1* and *Cluster 2* in t-SNE projection contain multiple subunits Figure 3.4A), the location of the samples within the clusters are, with few exceptions, closer to the samples from the same subunits.

This distinction is not as clear in most of the PCA biplots, due to the differences of the objectives of PCA and t-SNE transformation methods, i.e. PCA aims to preserve the global structure of the data whereas t-SNE reproduces the local structures. However, PC1 results show a clear distinction between SG3 samples from the other units, as SG3 samples are mainly clustered in PC1- and the majority of the SG1 and SG2 samples are located in PC1+ (Figure 3.7, Figure 3.10, Figure 3.11). PC1- elemental association is typical of the clastic silicate components and common immobile elements in basinal hydrothermal settings (e.g., Al_2O_3, TiO_2, SiO_2, Hf, Zr, REE, Th, Li, SrO, MgO, and TiO2) (Slezak et al., 2014, Olivo et al., 2018). SG1 samples from *Lapa Azul* and *Serra da Pindaiba* occurrences that plot on PC1- comprise the sandstone-siltstone lithofacies with a higher abundance of detrital grains (Fernandes et al., 2019a,b). On the other hand, PC1+ shows an association of mobile elements that are typical of the geochemical signature of pre-orogenic zinc

mineralization, and similar to the known deposits hosted in the carbonate sequence in the belt, such as Zn, Cd, Cu, Hg, In, V ± Sb, Se, Mo, Re (Slezak et al., 2014, Olivo et al., 2018; Fernandes et al., 2019a,b). It is commonly considered that mineralization events are captured by smaller principal components (Grunsky, 2010), as they represent localized processes. However, the fact that the metal association typical of hydrothermal activity in the Serra do Garrote Formation was captured in the first principal component, which corresponds to nearly 60% of the variance in the data, suggests that these elements are related to a widespread and/or, long-living hydrothermal activity without an efficient mechanism to focus the metal-bearing fluids, causing basin-scale sub-economical Zn and related metal enrichment (see Figure 3.7 for spatial distribution and Figure 3.10 for PCA biplot). This is consistent with the findings of Fernandes et al. (2019b), which document occurrences of zinc mineralization in the Serra do Garrote Formation throughout the basin.

Mineralized SGF samples with PC1+ have distinct PC2+ and PC2- elemental signatures (Figure 3.6, Figure 3.8, Figure 3.10), which could be explained by the distinct ore-related mineral compositions. Samples with higher zinc content (Figure 3.10) yielded PC2 values around -2 to +2. However, samples that show evidence of remobilization during orogenesis, characterized by the occurrence of Sph 3 (Cd-poor) along the main foliation (Figure 3.6 I-K, Table 3.3 and Figure 3.10 in Fernandes et al., 2019a), and lower zinc contents have higher loadings of PC +2.

Ponte Quemada samples are distinct from the rest of the SG1 samples both in t-SNE and PC plots. Intriguingly, element associations of PC2- include Cd, S, and organic carbon, whereas Zn plots with In and total carbon in the PC2+ field. Considering that Cd occurs commonly as a minor element within SGF sphalerite, this may indicate that Cd in the *Ponte Quemada* is also hosted in other mineral phases (Figure 3.10), probably related to distinct fluid-rock interaction processes.

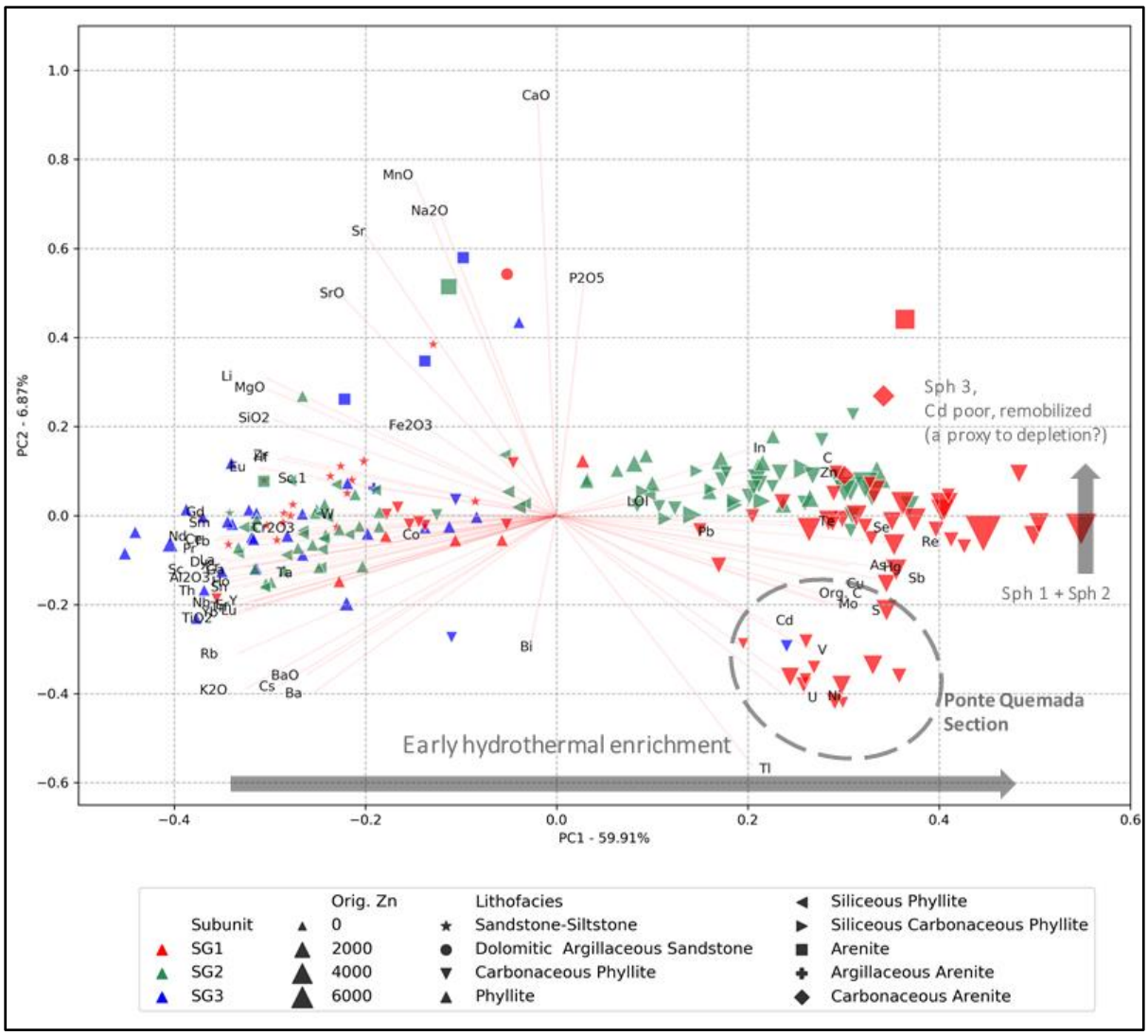


Figure 3.10 Interpreted PC1 vs PC2 biplot of SGF database. PC1- comprises element / oxide associations mostly related to carbonate minerals and / or detrital content. Note that in PC1+, Cd and Zn are separated slightly by the PC2 (see text for the interpretation). All of the SG3 samples are located on the negative side of the PC1 (PC1-) axis. The majority of the SG1 and SG2 samples are clustered in PC1+. Samples from Ponte Quemada occurrence (PC1+ and PC2- quadrant) are slightly shifted away from the rest of the group. There is a very close affinity between PC1 vs PC2 biplot and the major clusters revealed in the t-SNE projection (Figure 3.4A). While t-SNE emphasizes clusters better, PCA provides means of explaining these differences through their element / oxide loadings.

Similarly, to PC2, PC6 (Figure 3.11) separates two distinct elemental associations for the mineralized samples with PC1+. Samples with PC6- have an association of Zn, Cd, Cu and In which corresponds with typical composition of sphalerite and chalcopyrite inclusions, and higher

concentrations of Zn (note the triangle size). However samples with PC6+ represent an association of As, Sb, Te organic carbon, and S, which resembles the signature of sulfides (mainly pyrite) associated with organic carbon (Fe_2O_3 yielded PC6~0, as it occurs both in pyrite in the mineralized samples and in chlorite in the barren samples). Fernandes et al. (2019b) (in Figures 3.6 I-K) have

documented the occurrence of pyrite-rich bands where base metal inclusions were leached, and the PC6+ is tentatively interpreted to have captured the signature of this remobilization event.

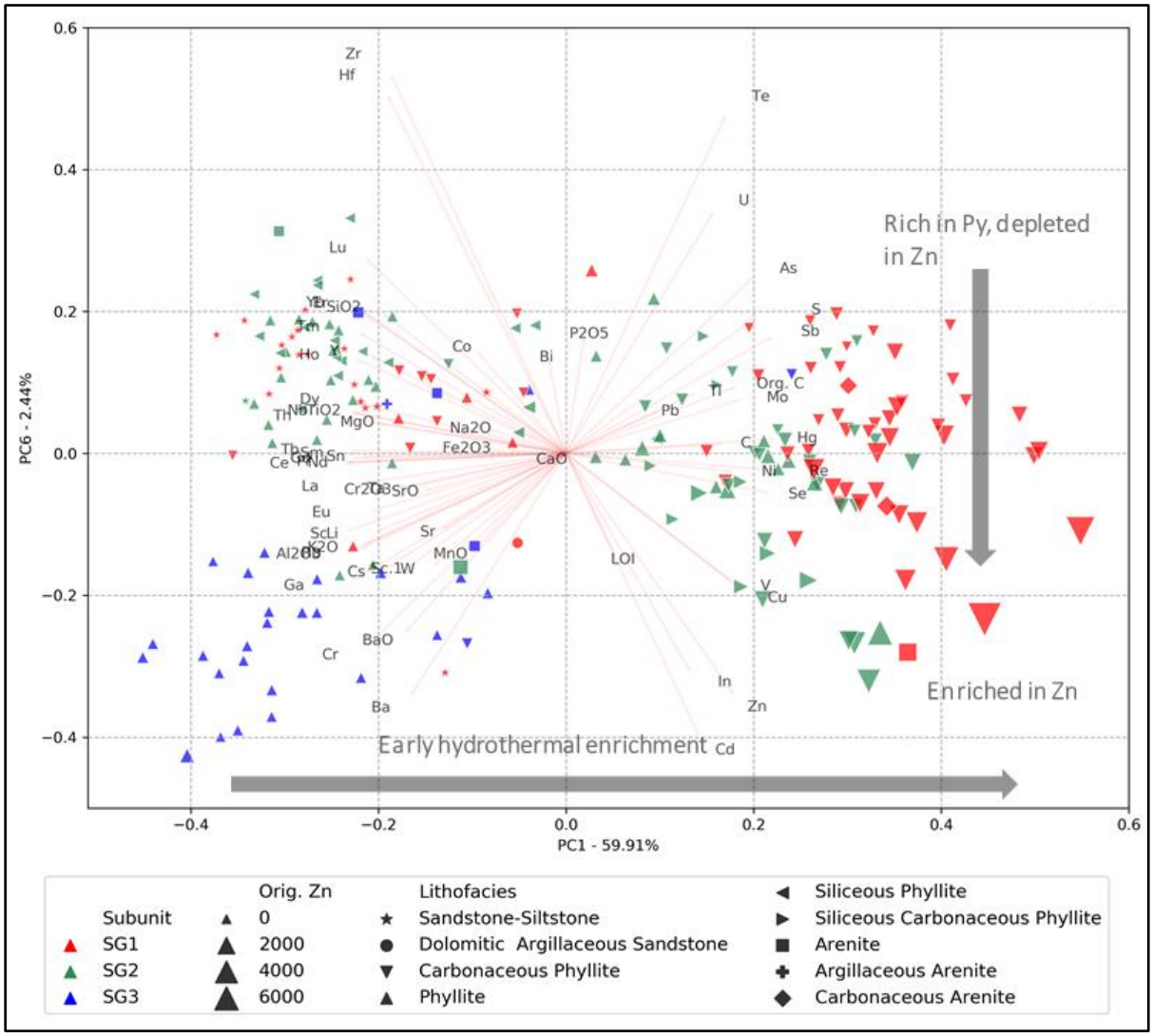


Figure 3.11 Interpreted PC1 vs PC6 biplot of the SGF database. Notice that, symbols of the samples show the lithofacies information. Rationale of the interpretations are presented in the text.

The relationship of PC3, PC4 and PC5 with mineralization (PC1+ and variation in Zn content) is not as clear as PC1, PC2 and PC6. A more detail petrographic and mineralogical study on the phyllite and carbonaceous phyllite samples from specific occurrences such as *Morro Agudo, Cafundo, North Extension, Olho D'Agua & Varginha, Mata Preta* and *Engenho Velho*, including

composition of silicate minerals and organic matter, would assist in better constraining the PC signature.

Implications for exploration

Interpretation of the principal components indicates that SGF has favorable geochemical signatures to be considered as a potential source rock for the known deposits located in the upper carbonate units. These signatures are characterized by: (a) a wide-spread, sub-economic enrichment in ore related elements, which is represented by PC1+, and is interpreted to be formed prior to the syn-orogenic event that formed the economic carbonate-hosted mineralization, and (b) evidence of depletion in some areas based mainly on PC2 and PC6 data (Table 3.2, Figure 3.10, Figure 3.11), which are interpreted to have occurred during the orogenic event as these show signature typical the mineralogical recorded in zones of metal remobilization (this study and Fernandes et al., 2019a). If indeed the SGF was one of the major sources of metals for the syn-orogenic economic carbonate-hosted base metal mineralization, then these signatures can be used for targeting carbonate-hosted base metal deposits by integrating with the spatial distribution. In this context, the most prospective areas are the ones in which there is evidence of both pre-orogenesis metal enrichment and syn-orogenic depletion. All sections with significant Pb-Zn resources in the carbonate rocks exhibit evidence of both enrichment and depletion in the SGF (Figure 3.7-Figure 3.9, Table 3.3) except Morro Agudo section, from which few samples were available. Curiously, Engenho Velho section has a similar signature, and yet, there is no deposit identified. Thus, based on this study *Engenho Velho* section is considered a priority target for exploration.

Table 3.2 Summary of the mineralization related principal components and their respective interpretation.

Component	Interpretation
PC1+	Ground preparation, enrichment in ore related elements
PC2+	Precipitation of Sph3, which is spatially associated with a depletion
PC6+	Depletion in ore related elements

Table 3.3 Summary table of the mineralization related events and their occurrences in the studied sections. Note that, except Morro Agudo section, all of the sections that have known mineralization have signature of all three principal components.

Section	PC1+	PC2+	PC6+	Known Deposit/Resource
Serra da Pindaiba		●	●	
Lapa Azul		●	●	
Ambrósia Norte	●	●	●	YES
Ponte Quemada	●		●	
Correco Rico	●	●		
Morro Agudo		●	●	YES
Cafundo	●	●	●	YES
North Extension	●	●	●	YES
Olho D'Agua			●	
Varginha	●	●	●	YES
Mata Preta		●	●	
Engenho Velho	●	●	●	**NO, potential to explore**

3.2.6 Conclusions

This is a pioneering exploratory study, which applied a multivariate analysis workflow to enhance the knowledge discovery from geochemical data, in order to provide insights about a mineralized belt. The novelty consists in the integration of Random Forest algorithm, t-SNE and PCA in geochemical data of rock units interpreted as potential sources for base metal mineralization in

order to identify zones with pre-enrichment and syn-ore depletions and assist in finding new targets for exploration. The Random Forest algorithm in unsupervised mode allowed to create a pair-wise distance matrix. t-SNE was then applied to embed this distance matrix to summarize the local structures of the data effectively in a 2D plot. t-SNE projection improved the understanding of the local structures and helped us to formulate directed questions which were explored through different PCs suggesting that non-linear manifold learning methods are potentially useful complementary tools in multivariate analysis of the geochemical data.

Principal component analysis allowed for the identification of multivariate patterns interpreted as the evidences for Serra do Garrote Formation to be considered as a source rock for the overlying deposits. PC1+ captured the signature of metal enrichment throughout the basin during the pre-orogenesis event. PC2 + and PC6+ are interpreted to reveal the signature of depletion during the orogenic event, which is coeval with the formation of the economic zinc mineralization in the carbonate rocks. All of the sections that have a known occurrence are represented by a combination of these events, except Morro Agudo which had limited samples. Interestingly, Engenho Velho section shares the similar signature with those sections with known deposit/resources, and yet, does not have any known deposit. Therefore, we suggest this location is a priority target for exploration. This innovative approach can be applied in other geological settings where the potential source rocks are exposed by drilling during regional or camp scale exploration in order to identify zones of enrichments and depletion to discover new exploration targets.

Acknowledgements

The authors would like to thank Dr. Neil Fernandes for the discussions on the geochemical rock sample database and also to Nexa Resources who granted collection and analysis of the rock samples. Dr. Ortiz acknowledges the support of the Natural Sciences and Engineering Council of

Canada (NSERC), funding reference number RGPIN-2017-04200 and RGPAS-2017-507956. Dr. Olivo also acknowledges the support of the NSERC Discovery Grant which supported analytical costs. Cevik would like to thank the Ministry of National Education Turkey for providing the scholarship funding of the M.A.Sc.

Data Availability

All of the data processing and analysis in this study are conducted by open source tools. The python/R code used to analyze the geochemical data is available as Jupyter notebooks at: https://github.com/geometatqueens

3.3. On the use of machine learning for mineral resource classification

3.3.1 Abstract

Mineral resource classification relies on the expert assessment of a Qualified Person to determine which blocks of a 3D mineral resource model are classified as measured, indicated, or inferred. The decision is often based on a combination of quantitative parameters related to the estimation process and qualitative decisions based on previous experience or preconceptions not captured in the numerical model. As such, the procedure is subject to inconsistency, that is, blocks with similar qualities may end up in different categories, mainly due to the subjective nature of the approach.

In this paper, we present a methodology to assist the qualified person in this task by clustering blocks with similar parameters and then classifying them into categories in a consistent and automatic manner that only requires the specification of a few hyper-parameters. The result is a consistent classification into resource categories comparable to the result of a classification done by a qualified person, but fully consistent and generated in a short time frame.

The procedure begins with repeatedly subsampling the block model to determine a distance matrix using an unsupervised random forest approach and then clustering the blocks using the associated distance matrix. Then, all the blocks in the model are classified using a supervised random forest approach, which gives a probability of belonging to each class. The initial class can be determined from the class probabilities. Support vector classification with a radial basis function kernel is utilized to smooth the boundaries between classes and define the final classification. An approach to tune the hyperparameters of smoothing is provided.

The methodology is demonstrated with two examples from two gold deposits. Results are comparable to the classification done by the project Qualified Person using conventional methods.

3.3.2 Introduction

Mineral resources are classified for public disclosure, based on their confidence level, into Inferred, Indicated, and Measured categories, and mining reserves into Probable and Proven categories (CIM, 2019; JORC, 2012; SAMREC, 2016). A reliable classification plays an important role in many downstream activities of a mining project since many parameters, such as ore tonnage and grades are used for planning and design purposes and depend on these classifications (Deutsch et al., 2007; Battalgazy and Madani, 2019). Furthermore, mineral resources and mining reserves classifications are required by financial institutions, investors and policymakers to make informed decisions such as determining royalties, regulation of taxations and making strategic decisions of investments (Emery et al., 2006). However, the task of classification of mineral resources and mining reserves is not trivial, and the determination of the criteria to outline these classes is a longstanding problem in the mineral industry (Deutsch et al., 2007). Important factors considered by the Qualified Person in charge of the classification comprise confidence in the geological model, spatial features of the deposit including the continuity of the grades and domains, the density of the informing data, the analytical quality control program and results, and the reliability of the estimation method (Stephenson and Stoker, 2001; Dohm, 2005; CIM, 2019). Some of these factors are difficult to capture by numerical models whereas other parameters are quantitative in nature and can be used to determine the confidence in the estimation method. Typical examples of variables considered during classification (these are called features in the machine learning literature) may include:

- The geological domain where the block is located;
- The kriging variance;
- Coefficient of variation of the estimate;

- The kriging pass;
- The distance to the nearest sample or drillhole;
- The average distance to the samples used in the estimation;
- The number of drillholes used in the estimation.

Although the international codes do not prescribe the methodology to define the classification, they suggest that the Qualified Person quantifies the confidence of the estimation to assign a class if it is warranted (CIM, 2019).

Some practitioners suggest using probabilistic approaches through geostatistical simulations (Dohm, 2005; Wawruch and Betzhold, 2005) or multi-gaussian kriging (Ortiz and Deutsch, 2003) whereas others claim that using a pure probabilistic approach is highly sensitive to the parameter selection and prone to mischief, and should only be used as supporting information to an approach based on geometric and geologic criteria, which is considered more transparent (Deutsch et al., 2007). Although the matter has been extensively studied and new proposals and comparisons of methods are recurrent in the literature (Emery et al., 2006; Silva and Boisvert, 2014), all the methods require a Qualified Person to define thresholds between classes, e.g. maximum distance to drill holes or maximum kriging variance to belong to a category. Therefore, it appears that the classification task will remain subjective regardless of the methodology used. However, one of the key requirements of any approach is its reproducibility so that the result can be easily audited (Coombes et al., 2014).

Common numerical approaches (based on geostatistical or geometric criteria) generate a block-by-block classification that often results in spatially inconsistent results that fail to respect the geological continuity and are impractical in terms of the mine plan. Several examples were

presented in Stephenson and Stoker (2001). Therefore, a smoothing procedure is usually needed as a post-process step (Deutsch et al., 2007; Stephenson et al., 2014).

The aim of this study is to provide a framework in which various types of data can be assimilated in a consistent and repeatable manner for resource estimation. The methodology section gives an overview of the proposed workflow and then explains each step in more detail. Two case studies are presented in section three and four respectively, and the results are discussed in the last section.

3.3.3 Methodology

Figure 3.12 provides a general overview of the steps involved in the proposed classification methodology. The starting point is a block model where each block is informed with a set of features as those described in the introduction (kriging variance, distance to nearest drillhole, number of samples used, etc.). All steps are done by using open source programming languages R and Python. The specific libraries are referred in the text in the corresponding sections.

1. **Preprocessing**: the first step of the workflow is to preprocess the block model data to remove unestimated blocks and label blocks that are classified as inferred by purely subjective expert criteria. This step also involves assigning qualitative and quantitative information to each block.
2. **Clustering**:
    2.1. **Similarity and distance metric**: a similarity measure is defined between the blocks, based on the underlying information available. Unsupervised Random Forest (RF) (Breiman, 2001; Shi and Horvath, 2006) is used to quantify the similarities between blocks which are then converted into distances.

2.2. **Clustering on subsets of blocks**: a k-medoids clustering algorithm (Schubert and Rousseeuw, 2019) is used to group blocks based on the precomputed distances. For computational efficiency, clustering is done on subsets of the whole block model.

2.3. **Classification of blocks into different categories**: the clusters defined in the subsets of blocks in the previous step are labelled with a resource category (i.e. Measured, Indicated, Inferred) based on the distribution of the features, and are used to train a Random Forest classifier (Breiman, 2001) to assign each block in the model into a category.

3. **Smoothing**: these initial classification results are then smoothed by using a Support Vector Machine (SVM) algorithm (Cortes and Vapnik, 1995). The degree of smoothing is controlled by two metrics, accuracy and percent change between classes. The algorithm can be applied several times to achieve the level of smoothing deemed adequate by the Qualified Person.

The following sections present the details of each step.

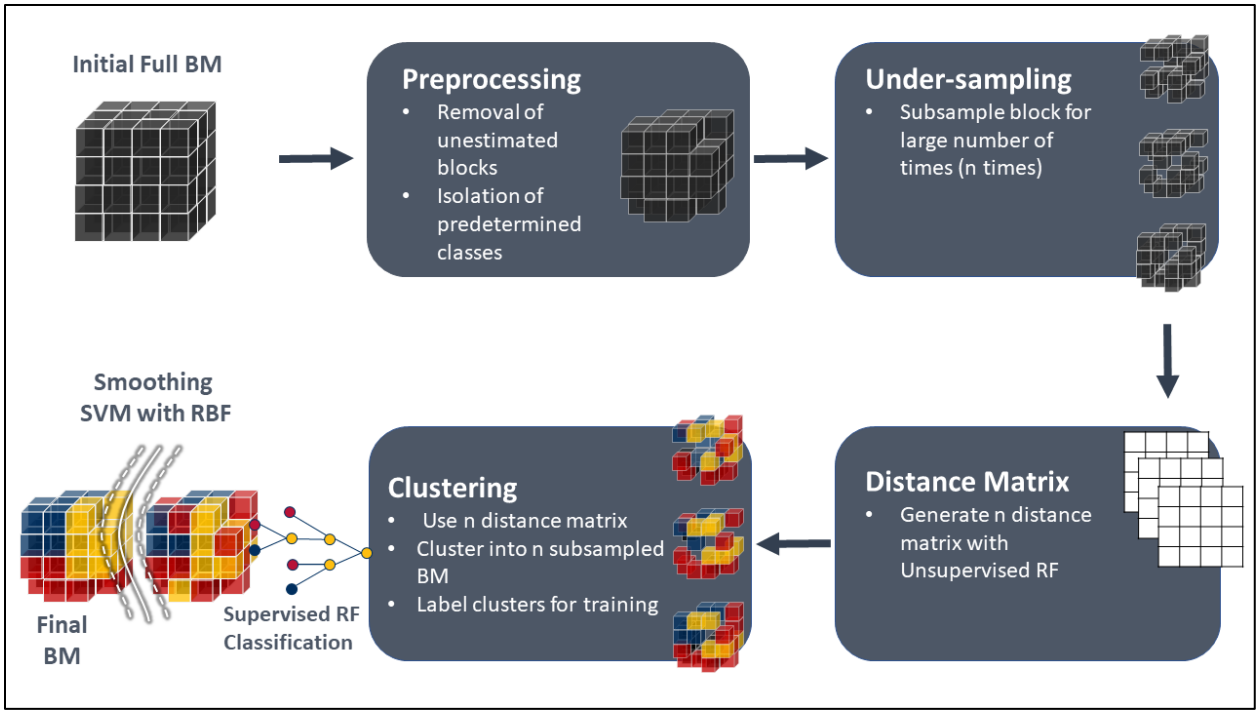


Figure 3.12 Workflow of the proposed approach. The initial block model (BM) is preprocessed to isolate blocks that will not be classified and assign categories by expert judgement. This is followed by sub-sampling of this preprocessed BM a large number of times (n times). By using Unsupervised RF, a pair-wise distance matrix is created for each sub-sampled BM. The distance matrices are then used to cluster the n sub-sampled block models with the k-medoids clustering method. Clusters are then labelled with corresponding resource classes. In the final step, supervised RF is used to classify the blocks in the entire BM by using these labelled sub-sampled block models as training information. Finally, SVM with a radial basis function (RBF) kernel is used to smooth the model and achieve the final classified block model.

Data Preprocessing

The blocks in the block model are defined as:

$$B_j = (x_{1j}, \dots, x_{Lj})$$

where $j = 1, \dots, N'$, and N' is the total number of blocks in the model, x_{lj} are the features recorded for each block, with L as the total number of features. The goal is to assign each block to a class k, where $k \in \{Measured, Indicated, Inferred\}$ or $k \in \{Indicated, Inferred\}$ if Measured is not considered a possibility by the Qualified Person:

$$B_j \leftarrow k$$

The practitioner needs to decide how many categories will be used in the subsequent steps: two, if the blocks are to be classified as inferred or indicated, or three if all three categories are to be used (inferred, indicated, and measured).

The pre-processing step often considers the following processes:

- Manual assignment of category to blocks based on expert judgement,
- Attribution of qualitative information for subsequent processing.

Block models generally contain a significant number of unestimated blocks, which may be due to estimation parameters imposed, lack of data, and/or domain specification. These blocks are removed from the block model in this step as they are unclassified for reporting purposes. Some of the blocks with specific properties may have a predetermined class. For example, mineral resources should be automatically excluded from classification or classified as inferred if there is lack of confidence in the geological model, since geological considerations always override any mathematical measure of uncertainty (Emery et al., 2006). Therefore, at this stage, specific classification labels (inferred or non-classified) are assigned to some blocks and these are excluded in the subsequent steps. The number of blocks remaining to be classified are N, that is $N' - N$ blocks were excluded or manually labelled.

The practitioner can also code and assign qualitative information to each block, such as the quality of the samples, confidence level in the geological model, or structural complexity of the geology in this step. There is no need for scaling or standardization of the data since the proposed approach is non-parametric and invariant to the scale of the data, as discussed next.

Clustering

Similarity and distance metric

In this step, the goal is to assign each block to a cluster of blocks with similar features. Any conventional clustering algorithm could serve this purpose as long as the distances or similarities between blocks in feature space are available. This may seem like a trivial task; however, depending on the estimation technique, each block may have parameters of very different nature, such as numeric (kriging variance, the average distance to composites, nearest drill hole, etc.), ordinal (quality of samples) or categorical (different estimation passes with levels of constrains), for which it is not straightforward to determine the distances between blocks in feature space. The Euclidian distance, for example, can be calculated just by considering kriging variance and average distance to composites. However, one cannot easily define the distance between blocks that are estimated in the first and second estimation pass. In addition, considering the qualitative features adds another layer of complexity.

In our application, we use the Random Forest (RF) approach (Breiman, 2001) to determine the similarity between blocks. RF is a machine learning method commonly used for supervised classification. It is an ensemble tree predictor which comprises a large number of decision trees. Each tree classifies the blocks in a randomly sampled subset (with replacement) of the original data, which is referred as *bagging* in the machine learning literature, and through a random selection of the features in each split (Breiman, 2001). The final classification can be achieved by majority voting from all the decision trees' results, or results can be reported as a probability inferred from the results of the ensemble of trees.

RF can also be used in unsupervised mode to create a similarity matrix between samples which then can be converted into a distance matrix (Shi and Horvath, 2006). The advantages of using RF to calculate the distance between samples is two-fold; firstly, RF is non-parametric and scale-invariant, that is, scaling the features to calculate the distance matrix is not required, and secondly, RF is robust to outliers (Shi and Horvath, 2006) and it handles categorical data (Breiman, 2001).

Calculating the pairwise distance matrix for each block in a block model that may contain thousands to millions of blocks is computationally expensive or in some cases impossible due to hardware limitations. In order to overcome the computational issues, we adopt a strategy based on sub-sampling the block model (down to a number of blocks, n_{sub}). The process is repeated a large number of times ($n_{realization}$) in order to reduce the information loss caused by using a subset of the original blocks to determine the distance matrix. Each realization based on a subset of the blocks yields a distance matrix, which is then used to cluster the corresponding subset of blocks as explained in the next section. The specific parameters used in our implementation are documented in Table 3.4.

In this study, $n_{sub} = 1000$ was chosen as it provides a reasonable computation time and a total number of 100 realizations were conducted ($n_{realization}$= 100), which ensures inferring a robust average behavior (Goovaerts, 1999).

The unsupervised RF method generates the distance matrix by using the following procedure. The original blocks are labeled as Class A and a synthetic dataset (henceforth labelled Class B) with the same number of features L and observations n_{sub} is created by independently sampling from the marginal distribution of each feature of Class A, creating a dataset where the features are uncorrelated (Figure 3.13).

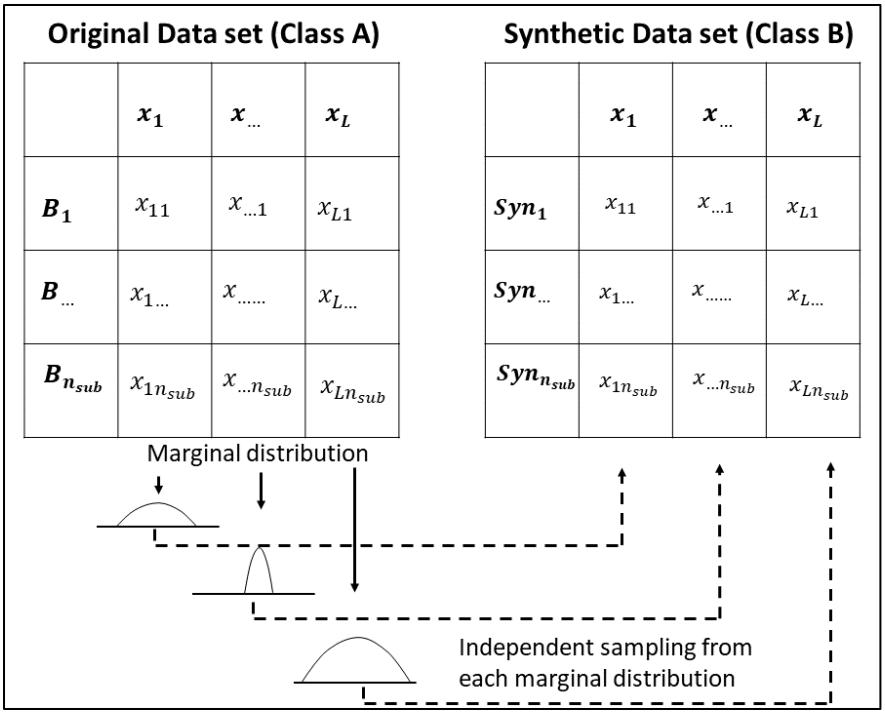


Figure 3.13 Schematic representation of synthetic dataset generation in unsupervised RF. Syn_i, i=1,…,n_{sub}, where i represents a synthetic record. The random sampling from the univariate distributions of the original variables or features removes the dependency of the variables in synthetic data set.

With these two data types (Class A and B), the problem becomes a supervised classification task and supervised RF is used as a classifier to learn to separate Class A and Class B. , In other words, supervised RF is used to separate the blocks from the original set, where the features are correlated to each other, from the synthetic set of blocks, where these features are uncorrelated. The similarity Sim_{ij} between two blocks B_i and B_j in the original set (Class A) is calculated by counting how often they end up in the same final tree node during training. Notice that blocks from the synthetic set (Class B) are disregarded, as our aim was to find the similarity between blocks in Class A. Blocks in Class B were only used to force the supervised random forest to assign blocks in Class A to different nodes when trying to separate two classes of blocks. At the end of the training process, these similarity counts are normalized by dividing by the number of trees in the forest,

which yields a symmetric positive-definite pairwise similarity matrix, and where each value lies in [0,1]. The similarity between blocks B_i and B_j, Sim_{ij}, can be expressed as follows:

$$Sim_{ij} = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} 1_{node_i = node_j}$$

where $1_{node_i = node_j}$ is the indicator function of $node_i = node_j$, $node_i$ and $node_j$ are the terminal nodes where B_i and B_j ended up in a tree of a forest. The dissimilarity between blocks B_i and B_j is defined as follows (Breiman, 2003; Shi and Horvath, 2006):

$$Dissim_{ij} = \sqrt{1 - Sim_{ij}}$$

The dissimilarity is used as a distance metric for subsequent steps.

There are a few hyper-parameters to be specified by the practitioner to use unsupervised RF, e.g. number of trees in the forest, n_{tree}, number of features to be considered in each split decision, m_{try}, and number of forests to average resultant distance matrix to achieve robust dissimilarities, n_{forest}. Breiman (2001) showed that the generalization error of the forest always converges as the number of trees increase, i.e. there is an upper limit of the number of trees to increase the performance of the forest. In the supervised case, where the class labels are known, this limit can be found by cross-validation. In our case, since we cast this as an unsupervised problem, we chose a large number to ensure this convergence, i.e. $n_{tree} = 1000$ as depicted in Table 3.4. Shi and Horvarth (2006) presented a sensitivity analysis on proximity estimation with unsupervised random forest. In their study, it is shown that proximities are not highly sensitive to the number of forests as long as there are at least 5 forests to be averaged. The quality of the clusters, for which the rand index was used as a performance metric, reaches a plateau after total of 5 forests.

Therefore, $n_{forest} = 5$ was used in this study (Table 3.4). The default value, the square root of the total features, was used for m_{try}, since both Breiman (2001) and Shi and Horvarth (2006) report that RF is not sensitive to the choice of this hyper-parameter as long as it is not chosen close to low or high extremes, e.g. only one feature or maximum number of features.

Clustering of subsets of blocks

Each of the resultant distance matrices ($n_{realization}$) are used to run a conventional clustering algorithm called Partitioning Around Medoids (PAM) or simply k-medoids (Schubert and Rousseeuw, 2019). K-medoids is chosen because it accepts a precomputed distance matrix as an input, but any other clustering algorithm that accepts the precomputed distances as input could be used as well.

K-medoids aims to partition the data into a predefined k number of clusters (in our case, 2 or 3) and minimizes the within-cluster distances or dissimilarities. It randomly chooses k representative blocks, so-called medoids, and assigns each sample to the closest medoid. Then, recomputes the new representative samples medoids to minimize the within-cluster distance. This objective can be expressed as (Schubert and Rousseeuw, 2019):

$$Total\ Deviation = \sum_{i=1}^{k} \sum_{B \in C_i} Dissim\,(B_j, B_{m_i})$$

where $B \in C_i$ refers to each block in the ith cluster C_i and m_i refers to the medoid of the corresponding cluster. Note that the clusters generated in this step are labelled with an identifier that is not yet linked to a mineral resource category and need to be labelled with a resource class in order to be used as training data for classification of the block model.

The labeling task is automated and based on the premise that:

- Each feature that is used for clustering should be chosen and therefore can be considered as a proxy for confidence in estimation,
- Hence, clusters identified in this step should represent different confidence levels in estimation, i.e., it is expected that blocks are grouped into different clusters since they have different confidence levels in estimation.

Therefore, it can be reasonably expected that the cluster with highest confidence level in estimation should have more favorable average values for the block features, e.g. lower average distance to samples and/or holes, lower average distance to nearest 3 holes or lower estimation variance.

Based on this premise, each cluster is labeled into a resource class with the following criteria;

1. Label a given cluster as "*measured*" (or "*indicated*" when k = 2), if the mean value of the "average distance to composites" **AND** the mean value of the "estimation variance" of cluster are the lowest (most favorable),
2. a. If k = 2, label the remaining cluster as "inferred",

   b. If k = 3, label a given cluster as "inferred" if the mean value of the "average distance to composites" **AND** the mean value of the "estimation variance" are the highest (least favorable), and label the remaining cluster as "*indicated*",
3. Label the clusters in a given realization as "*ambiguous*" if the above conditions are not satisfied, consider these clusters as "not qualified" and do not use these clusters as training data.

As seen from the previous description, blocks used for training are those where we are most confident about their assignment to a class.

The result of this process is a set of multiple realizations of training data with their corresponding training labels that will be used for classification as explained in Section 2.2.3.

Classification of blocks into category

The entire block model (excluding the blocks left out during the pre-processing) is, then, classified into these resource classes by using RF in supervised mode. The process is repeated $n_{realization}$ times, as each subset of the block model provides a different training set. The result is $n_{realization}$ classifications with different class labels for each block. The final class of each block is assigned based on the majority vote of this ensemble or averaging the class probabilities of each classification step. Both approaches yield similar results.

The same RF hyperparameter values used in the unsupervised setting are used in this case, e.g. m_{try} is set to the square root of the total number of features and n_{tree} is set to 1000 (Table 3.4) and no pruning was applied.

Smoothing

Classification on a block-by-block basis may yield classes that are not spatially contiguous. This is not desired in terms of mining, mainly for operational reasons, considering that these classes will inform the decision of classification of reserves and may have a downstream effect on production. Therefore, a post-processing is required to generate spatially contiguous classes from the initial classification. Notice that spatial continuity is not explicitly imposed in the classification approach, although a degree of spatial continuity is naturally expected due to the autocorrelation of the input features. An alternative to imposing the spatial continuity as a post-process is to directly impose it during clustering. Romary et al., 2012, presents a method that uses hierarchical

clustering with spatial constraints for geological domains. In this study, we applied another machine learning method to control the smoothness of the resulting classification.

There is a trade-off between maximizing the spatial continuity and minimizing the within-cluster differences. The continuity was established by smoothing the initial class in a way that the practitioner can control the number of blocks for which the final class after smoothing is changed, as well as the change in the total volume of the classes. Two metrics used to control this are explained later in this section.

The proposed approach casts this task as a classification problem and uses the Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel to reclassify blocks just by using their spatial information, i.e. their coordinates in 3D space. SVM fits in the smoothing context well since the SVMs' objective is to define the widest soft margins between classes, by allowing to make some errors in classification of the blocks within this margin which gives an opportunity to generate a smooth boundary. SVM aims to maximize the margin, M, between classes by allowing the user to tune the tolerance for making errors with a cost parameter C (James et al., 2013). Friedman et al. (2001) show that larger values of C will discourage misclassifications, as the boundaries between classes will be more rugged, whereas lower values of C will cause smoother boundaries.

Nonlinear boundaries can be inferred by projecting the original data into higher dimensional space and drawing linear boundaries in this space, which in turn generate non-linear boundaries in the original features space. However, projecting the data to a higher space can become unmanageable when the data becomes very large and/or the projected space gets very high in dimension. However, SVM simplifies the computation by using inner product of the observations rather than

projecting them to the higher dimensional space explicitly. The inner product is further simplified by approximating the distance between the observations with a kernel. We used *a radial basis function (RBF)* kernel since it has a local behavior, allowing nearby samples have a much higher weight on the classification boundary (James et al., 2013).

The RBF kernel depends on a parameter γ that scales the distance between samples (higher γ values lower the radius of influence of each observation). In order to find optimum C and γ values, the algorithm is trained and applied to the entire block model repeatedly, providing X, Y, and Z coordinates as inputs and initial classes as labels, varying both parameters C and γ to map the accuracy (percentage of the re-classified blocks to smooth the block model) and the change in total number of blocks in each class. The practitioner can control the trade-off between consistency of the classes and the spatial continuity, by selecting the C and γ values that maintain a reasonable accuracy, while allowing a controlled change in the classification of the blocks. The result of applying SVM with the selected C and γ parameters is the smoothed classification of the block model. This process can be repeated to increase the smoothing, if it is deemed as required by the Qualified Person.

The complete algorithm and parameters for the proposed approach are shown in Table 3.4.

Table 3.4 Algorithm for the proposed approach.

	Steps	Description	Objective	Hyper-parameters
Initial Class	1	**Input:** Preprocessed block model with estimation parameters		
	2	**Loop starts:**	Achieve robust statistics	$n_{realization}$ = 100
	2.1	***Under sample block model***	Alleviate computational burden	n_{sub}= 1000
	2.2	***Unsupervised RF***	Get distance matrix	n_{forest} = 5 n_{tree}= 1000 m_{try} = sqrt(features)
	2.2	***Clustering - PAM***	Identify clusters with distinct confidence level	k = 2 (or 3 when necessary)
	2.3	***Label clusters*** **if** k = 2; **if** all numeric variables have most favorable values in average; label as "indicated", label the other cluster as "inferred" **else if** flag the realization and do not use in training **if** k = 3; **if** all numeric variables have most favorable values in average; label as "measured", **AND** **if** all numeric variables have least favorable values in average; label as "inferred", and label the remaining cluster as "indicated" **else if** flag the realization and do not use in training	Map clusters to resource categories in subblocks	
	2.4	***Supervised RF***	Map resource categories of subblocks to entire block model	n_{tree}= 1000 m_{try} = sqrt(features)
	2.5	**Loop ends**		
Smoothing	3	Average class probabilities of $n_{realization}$ block models, **OR** count votes of $n_{realization}$ class for each block and use majority vote for classification of given block	Define resource classes	
	4	Extract a slice of the block model	Alleviate computational burden	Variable thickness
	5	Grid search for SVM with RBF parameters for C between 10^{-3} and 10^{3}: for γ between 10^{-3} and 10^{3} Use X, Y, Z for training, predict classes Return C and γ values which yields desired accuracy and percent change values	Selection of smoothing parameters	Accuracy and percent change
	6 6.1 6.2	Loop starts: While stopping criteria is not achieved; Run SVM with RBF on full block model; return accuracy & percent change **Stop** when accuracy and percent change is converged or when another user specified stopping criteria is achieved	Smoothing	C and γ; depends on the results of **step 5** **Stopping criteria;** such as min accuracy, max percent change, max smoothing pass etc.

3.3.4. Case Study I

Description and available information

The data used in this study is an estimated block model of a shear hosted lode gold deposit in West Africa with an ordinary kriging estimation method. For each block the following parameters, that are typical to a kriging estimation, are available:

- Estimation variance (numeric);
- Number of composites used in estimation (numeric);
- Number of drill holes used in estimation (numeric);
- Average distance to composites (numeric);
- Local block covariance (numeric);
- Estimation pass (ordinal categorical); and
- Coordinates of the centroids of the blocks (X, Y, Z in meters).

There are four estimation domains for which 419,044 blocks are available. Blocks are 10m x 10m x 3m in size. The exploratory data analysis revealed that a few of the blocks (n = 614) were not estimated. These blocks are removed from the analysis, making the total number of blocks to be classified equal to $N = 418{,}430$.

A summary table of the parameters with their descriptive statistics is given in Table 3.5. The counts for the different kriging passes are shown in Table 3.6.

Table 3.5 Descriptive statistics of the numerical estimation parameters of Case Study I.

		Count	Mean	Median	Std. Dev.	Minimum	Maximum
Coordinates	**X**	418,430	40,047	40,045	157	39,415	40,395
	Y	418,430	19,694	19,765	469	18,805	20,595

	Z	418,430	670	730	253	22	1,000
Estimation Parameters	**Est. Var.**	418,430	0.29	0.29	0.15	0.03	1.00
	Avg. Dist. to Comp.	418,430	37.23	26.25	31.55	2.83	203.76
	Local Block Cov.	418,430	0.29	0.28	0.04	0.21	0.48
	Number of Holes	418,430	3.53	3	2.06	1	16
	Number of Composites	418,430	14.95	16	2.71	2	16

Table 3.6 Number of blocks estimated in each estimation pass

Estimation Pass	Count
1	292,340
2	3,809
3	88,563
4	33,718

Based on the experience of the Qualified Person, all blocks are to be classified as indicated or inferred. Therefore, there is no preprocessing requirement except the removal of the blocks that are not estimated, as previously mentioned. The number of categories is $k = 2$.

Classification with machine learning

The proposed methodology is applied by using the 5 estimation parameters shown in Table 3.5.

The clustering and initial classification of the block model involves the following steps:

1. Random Forest (RF) implementation in the statistical programming language R (Liaw and Wiener, 2002) and modified functions in Shi and Hovard (2006) are used to extract the distance matrices. Hyper-parameters are n_{sub} = 1000, $n_{realization}$ = 100, m_{try} = 2 (closest integer to square root of the total features), n_{forest} = 5,
2. This distance matrix is then clustered into two confidence groups by using partitioning around medoids (PAM) algorithm implemented in the R programming language (Maechler et al., 2019). The only hyper-parameter, number of clusters, is set to $k = 2$,

3. Resultant clusters are labeled with a resource category (as explained previously),
4. Supervised RF is used to classify all blocks into these two categories. The result is the total of 100 ($n_{realization}$) block models with two new variables: class labels and class probabilities.
5. These 100 realizations are then merged into one block model; class labels are merged by counting the majority vote, and probabilities are merged by averaging. Initial classes, i.e. classes before smoothing, are achieved at this point.

The Python implementation of support vector machines with radial basis kernel is used to conduct smoothing operation. A grid search in logarithmic scale was conducted to choose optimum C and γ parameters to meet the smoothing criteria. It is usually sufficient to search for a logarithmic grid from 10^{-3} to 10^{3} (Pedregosa et al., 2011). These values were adjusted after the first pass of search, in a way that optimum values of the two metrics of interest, accuracy and percent change, is covered (see Figure 3.14 for more detail). A combination of 6 parameters of C (between 10^{-3} – 10^{2}), and 7 parameters of γ (10^{-7} – 10^{-1}) were tested. This grid search was done on a 50m thick slice of the deposit to have a reasonable computation time.

The smoothing criteria are as follows:

- Allow 10% of the blocks to be reclassified differently, i.e. *accuracy* should be around 90%.
- Choose the minimum change in total volumes in classes, i.e. reclassified blocks should not favor a class heavily; *Percent change* close to zero.

The best combination of C and γ was determined in this case was $C = 10^{-1}$, $\gamma = 10^{-5}$. As it is shown in Figure 3.14, these parameters are best in terms of our criteria; accuracy is close to 90% and percent change closest to 0%. The SVM is then used to smooth the entire block model with these

parameters. C = 0.01 and γ= 0.0001 also provide good results in terms of accuracy (~90%), however, the percent change shows that, one class is slightly more penalized than the other.

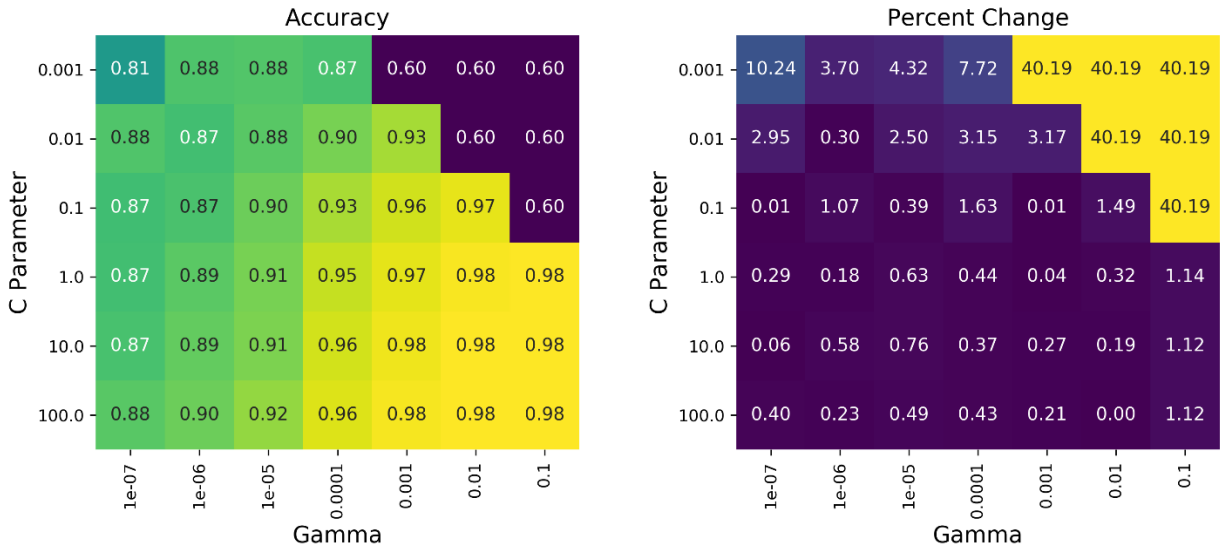


Figure 3.14 Grid search for accuracy and percent change in SVM classification with RBF kernel. Note that, target thresholds, 90% and 0% percent change are located approximately in the center of the grid to ensure the local optimum is captured with these parameters.

Results and Discussion

The quality of the clusters was determined by comparing the distribution of the estimation parameters in different classes with a qualified person (QP) classification and through a visual assessment of the spatial continuity.

As a result of applying the Unsupervised RF approach, a total of 78% and 22% of the blocks are classified as indicated and inferred, respectively. Figure 3.15 shows that almost all indicated blocks are selected from estimation pass 1, but not all pass 1 block are grouped in the indicated class which suggest that classification is not only driven by the estimation passes, but also considers other parameters.

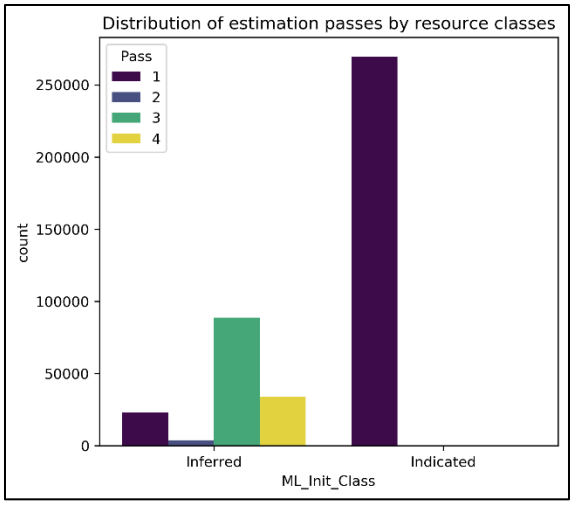


Figure 3.15 Distribution of the estimation passes based on classification results.

Most of the blocks that are classified as indicated have a lower average distance to composites, which is considered as one of the proxies of a confident classification (Figure 3.16). Boxplots show that the majority of the blocks in the indicated class are well separated from the blocks in the inferred class. Similarly, the estimation variance is also considerably lower in indicated blocks as compared to inferred blocks. In addition, it can be seen that the number of drill holes and composites used in estimation rarely reaches values lower than 3 and 16, respectively, in the class of indicated blocks, whereas a considerable number of blocks are estimated with lower number of drill holes and composites in the inferred class.

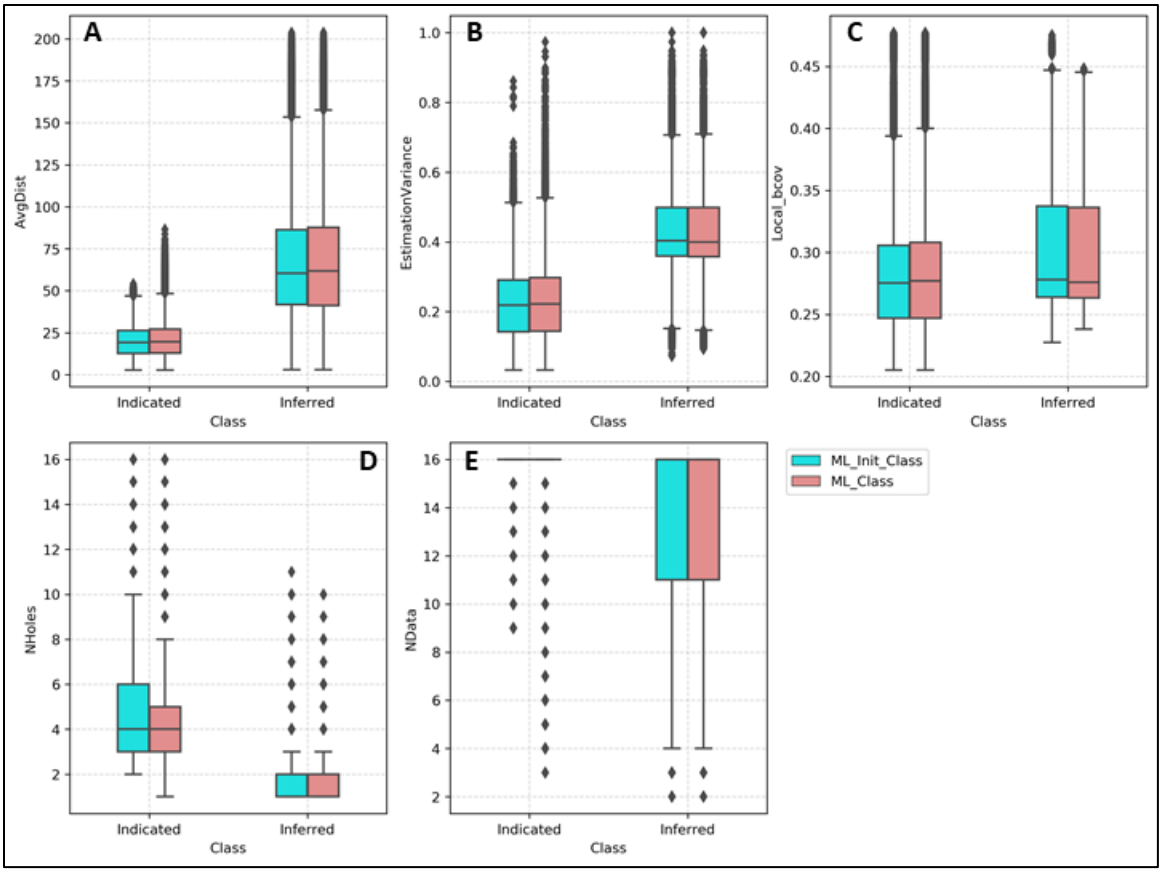


Figure 3.16 Comparison of some of the estimation parameters between classes. Distributions are; A) Average distance to informing samples (composites), B) estimation variance, C) local block covariance, D) number of informing drill holes, E) number of informing samples.

A visual comparison of the estimation parameters, initial class and the smoothed class is presented in Figure 3.17. Final classes are spatially continuous and respect the estimation parameters. Ideally, results would be compared with a classification done by a Qualified Person which was not available at the time of the study (See Case Study II for such comparison).

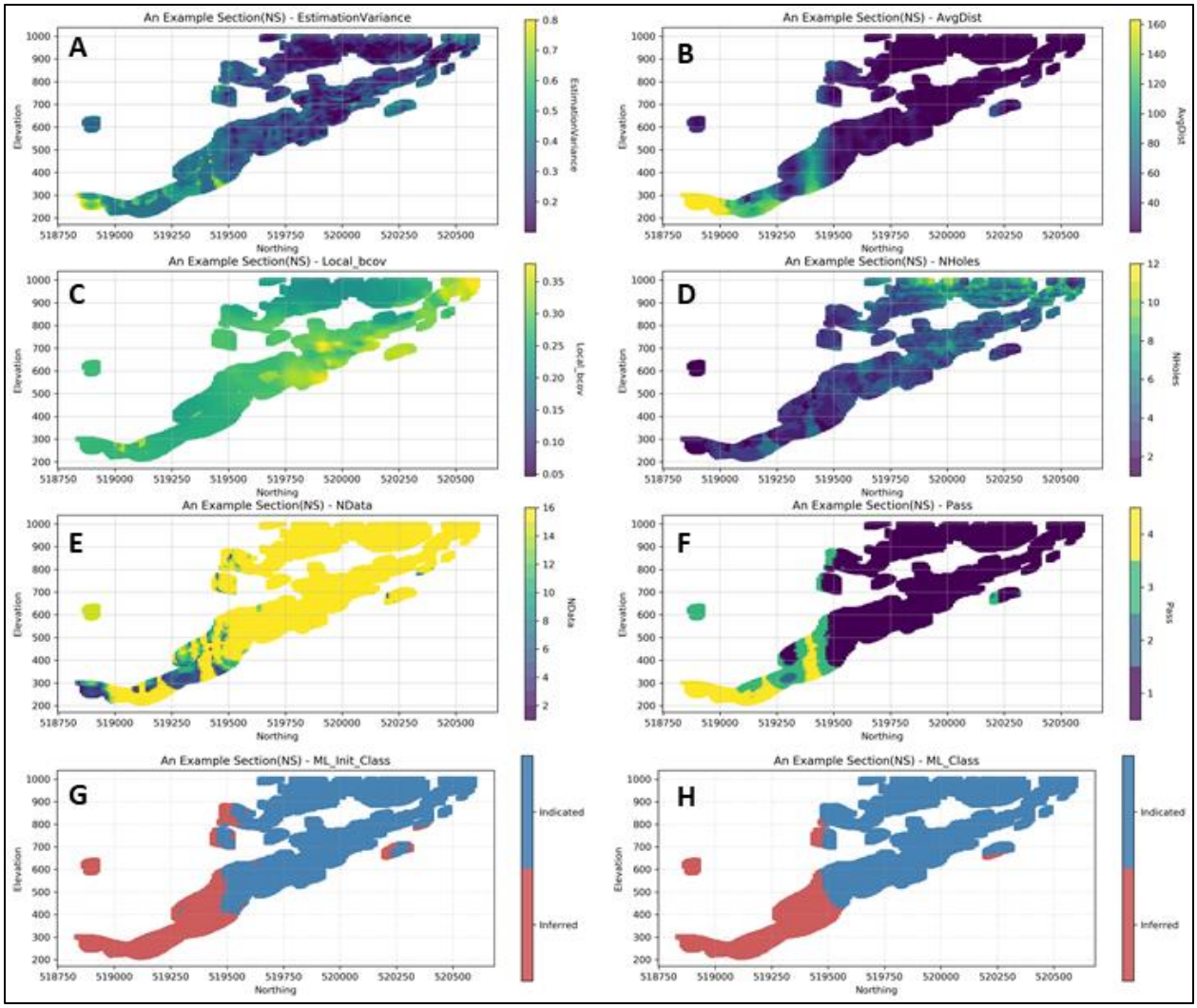


Figure 3.17 A representative section of the estimation parameters that are used for classification and the classification results. A) average distance to composites, B) estimation variance, c) local block covariance, D) number of composites, E) number of drill holes, F) estimation pass, G) initial class (pre-smoothing) and H) final class (after smoothing).

3.3.5. Case Study II

Description and available information

The second case study is a gold project in South America. The full block model has around 18 million blocks of which only 2.9 million were estimated. The blocks that are not estimated were removed from the analysis.

In this case, in addition to the estimation parameters that were used in the previous case, two geometric parameters are also provided for each block:

- A categorical variable, *Init Class*, is provided which represents three categories:
    - Category 2 – blocks that have 3 holes within 40m;
    - Category 3 – blocks that have 3 holes within 70m; and
    - Category 0 – none of the above. It is also predefined that, category 0, should be classified as inferred.
- *Dist. class* is defined as the average distance to the closest 3 holes.

A summary table of the all parameters used in this classification is given in Table 3.7. The counts of the categorical variables Kriging Pass and Geometric Class are provided in Table 3.8.

Table 3.7 Descriptive statistics of the numerical estimation parameters of Case Study II.

		Count	Mean	Median	Std. Dev.	Minimum	Maximum
Coordinates	**X**	115,400	678,662.79	678,590.36	350.99	677,806.51	679,445.22
	Y	115,400	544,067.74	544,179.48	503.73	543,101.86	545,481.99
	Z	115,400	202.06	207.00	126.49	(257.00)	431.00
Estimation Parameters	**Avg. Dist. to Comp.**	115,400	39.47	38.18	11.73	4.27	106.70
	Number of Holes	115,400	3.82	4	1.03	1	7
	Number of Composites	115,400	11.18	12	2.10	1	15
	Est. Var.	115,400	0.42	0.43	0.20	0.00	1.45
	Distance to 3 holes	115,400	27.78	26.03	9.99	1.61	69.36

Table 3.8 Total blocks per kriging pass and geometric class (Init Class).

Kriging Pass	Count
1	109699
2	5701

Init Class	Count
2	79510
3	35890

Classification with machine learning

The same settings that were used in Case Study I are used to conduct the unsupervised clustering task (see Table 3.4). Clustering results are labeled and used as training data. The entire block model is then classified. The smoothing operation is conducted in the same manner as well. The grid search results for the hyper-parameters of SVM with RBF are presented below (Figure 3.18). C = 0.1 and γ= 0.0001 are chosen since they yield an accuracy value of 90% and the percent change value of around 2.2% that matches the criteria defined in this study.

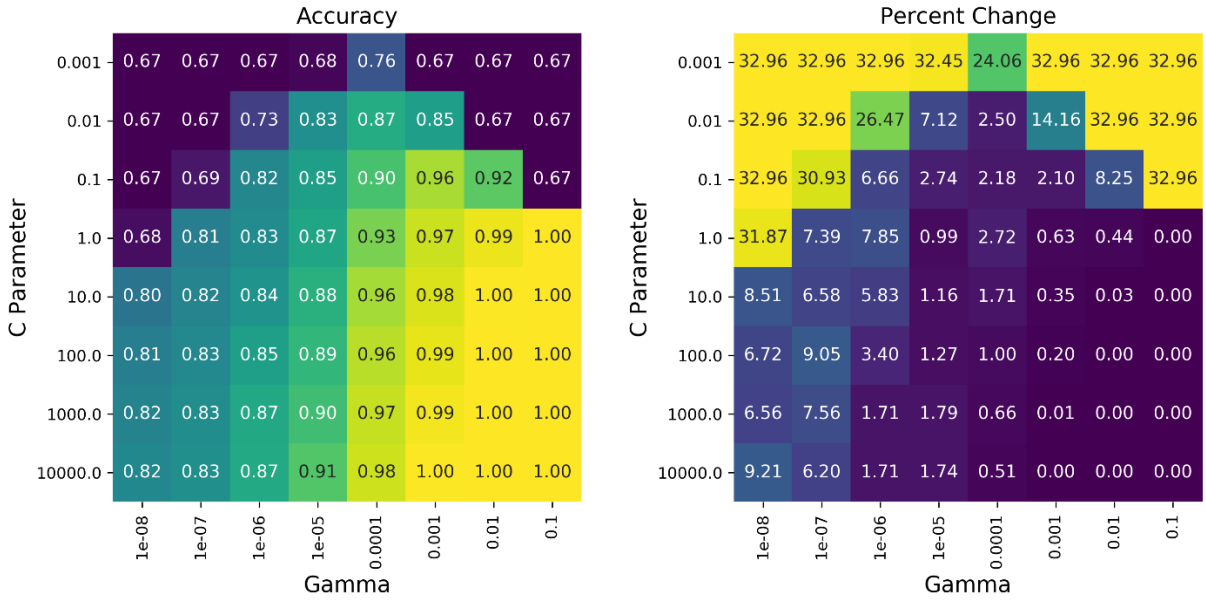


Figure 3.18 Grid search for accuracy and percent change in SVM classification with RBF kernel. Note that, target thresholds, ~90% and ~0% percent change are located approximately in the center of the grid to ensure the local optimum is captured with these parameters. C = 0.1 and gamma = 0.0001 was chosen in this instance.

Results and discussion

Comparison with an QP classification was made as follows:

- By using confusion matrices to quantitatively assess the degree of agreement between two approaches (Figure 3.19),

- By comparing the distribution of quantitative parameters through (Figure 3.20),
- Visually through cross sections, merely as a qualitative assessment (Figure 3.21, Figure 3.22),
- By presenting the average silhouette scores of the classes to quantitatively assess the classification quality (Table 3.9).

The silhouette score is a metric used to quantify the quality of clusters in terms of distances between members in the same cluster and the distances between members in different clusters. Each sample in a cluster can have a silhouette score between [-1,1]. A silhouette score close to 1 means that the sample is very close to the other members in the same cluster and far from the samples from another cluster, while a score close to -1 means that the sample is located in the wrong cluster. A score close to zero means that samples are located close to the boundary between clusters. The average silhouette score is used to assess the quality of the clusters (Rousseeuw, 1987).

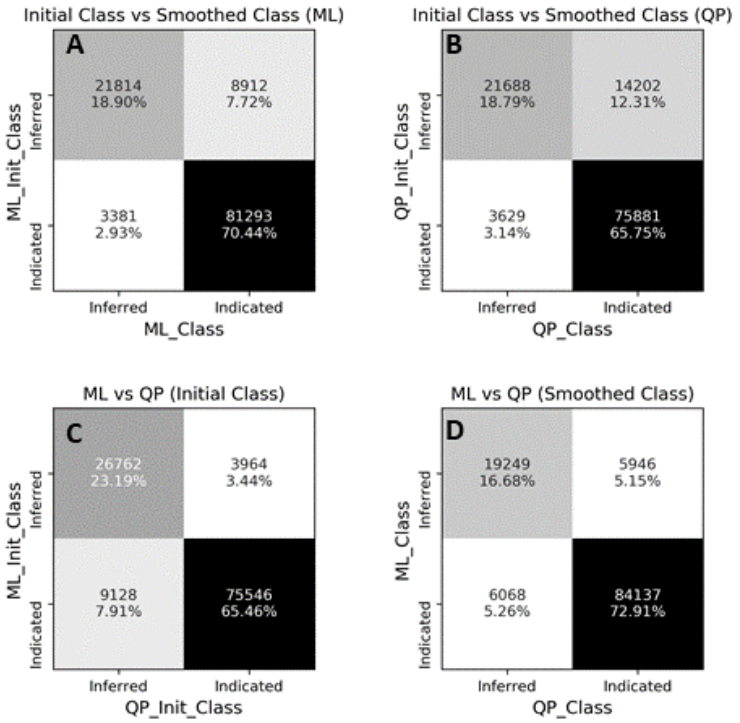


Figure 3.19 Confusion matrix between QP classification and the machine learning classification (ML Class). A) Comparison between initial (before smoothing) and smoothed classification results by machine learning (ML) approach, B) initial and smoothed results for classification done by a qualified person (QP), C) comparison of initial results, and D) smoothed versions of ML and QP classification approaches.

The confusion matrix between QP and machine learning classification shows that the two approaches have an agreement on approximately 90% of the blocks, of which near 73% are indicated blocks (Figure 3.19, D). It also shows that QP allowed approximately 15% of the blocks to be re-classified during smoothing, whereas the ML smoothing approach proposed in this study controlled the smoothing at approximately 10% (Figure 3.19, A-B).

Boxplots also show that the distribution of the estimation parameters between the two approaches are comparable and present clear separation between classes; except for the estimation variance and kriging efficiency, which is a derivative of the estimation variance (Figure 3.20).

Silhouette scores of the two approaches are also close (Table 3.9), having the ML approach slightly higher values, i.e., better separation between classes. However, it should be noted that, the silhouette score, or any other metric that can be applied to evaluate clustering performance by using distances, cannot be considered as a sole performance evaluator, but can be used as a proxy to the performance. This is because, besides the debates about their validity in different cluster shapes which is not the scope of this study, the silhouette score or similar metric only consider numeric variables unless a pre-computed distance matrix is provided. In contrast, classification of the blocks considers categorical (or ordinal) variables as well.

In this case, we do not have access to a distance matrix between all the blocks, due to computational issues that arise for large models, instead we only have access to the distance matrices between subsets of blocks.

Table 3.9 Comparison of the average silhouette scores of the two approaches, before and after smoothing. Data is scaled by subtracting the mean and dividing by the standard deviation.

	Avg. Silhouette Score
Final ML Class	0.32
Initial ML Class	0.34
Final QP Class	0.29
Initial QP Class	0.27

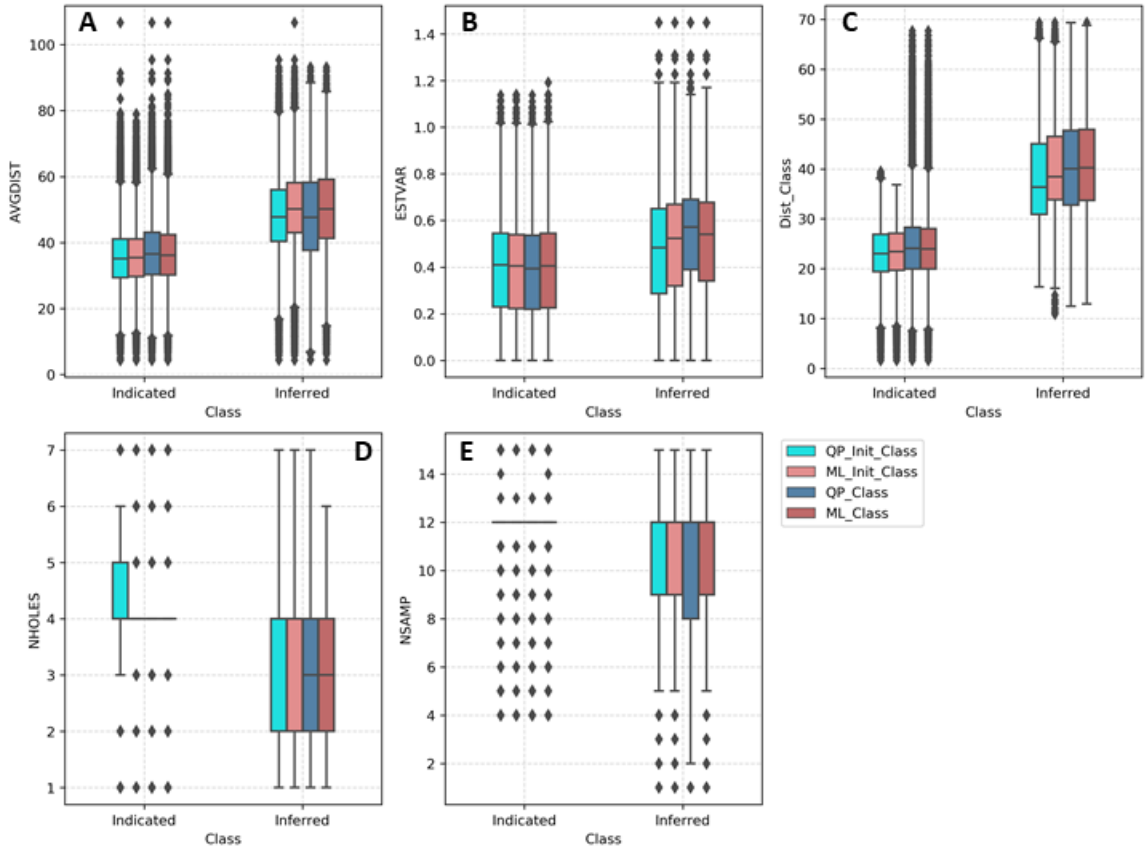


Figure 3.20 Boxplots that show the distribution of the estimation parameters among classes by different classification approaches. Abbreviations are; QP: Qualified person, Init. Class: Initial class (before smoothing, ML: Machine learning, QP and ML class: final classes after smoothing. Distributions are; A) Average distance to informing samples (composites), B) estimation variance, C) average distance to nearest 3 drill holes, D) number of informing drill holes, E) number of informing samples.

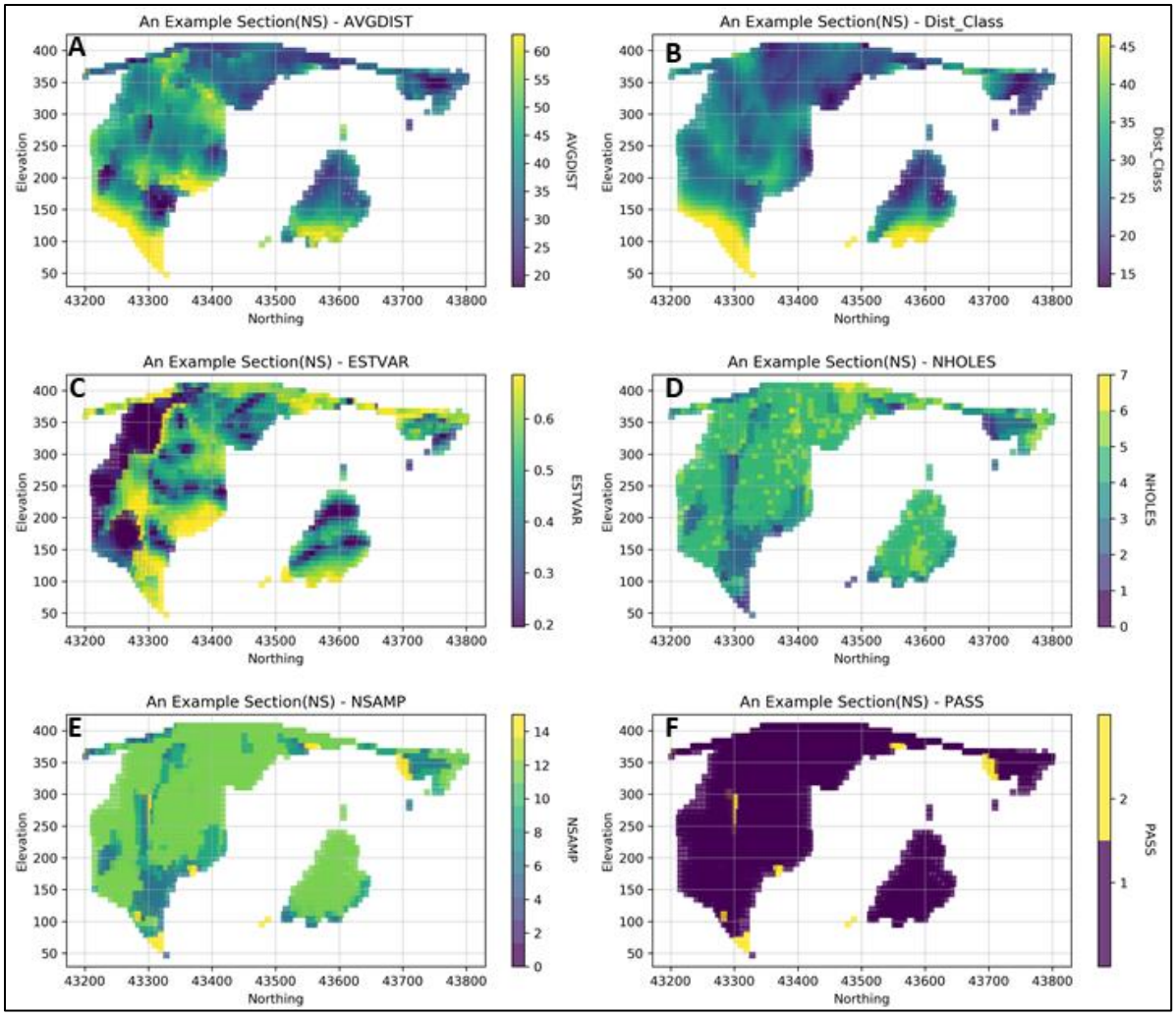


Figure 3.21 Estimation parameters for Case Study II. A) Average distance to composites, B) average distance to closest 3 drill holes, C) estimation variance, D) number of informing drill holes, E) number of composites, F) estimation pass.

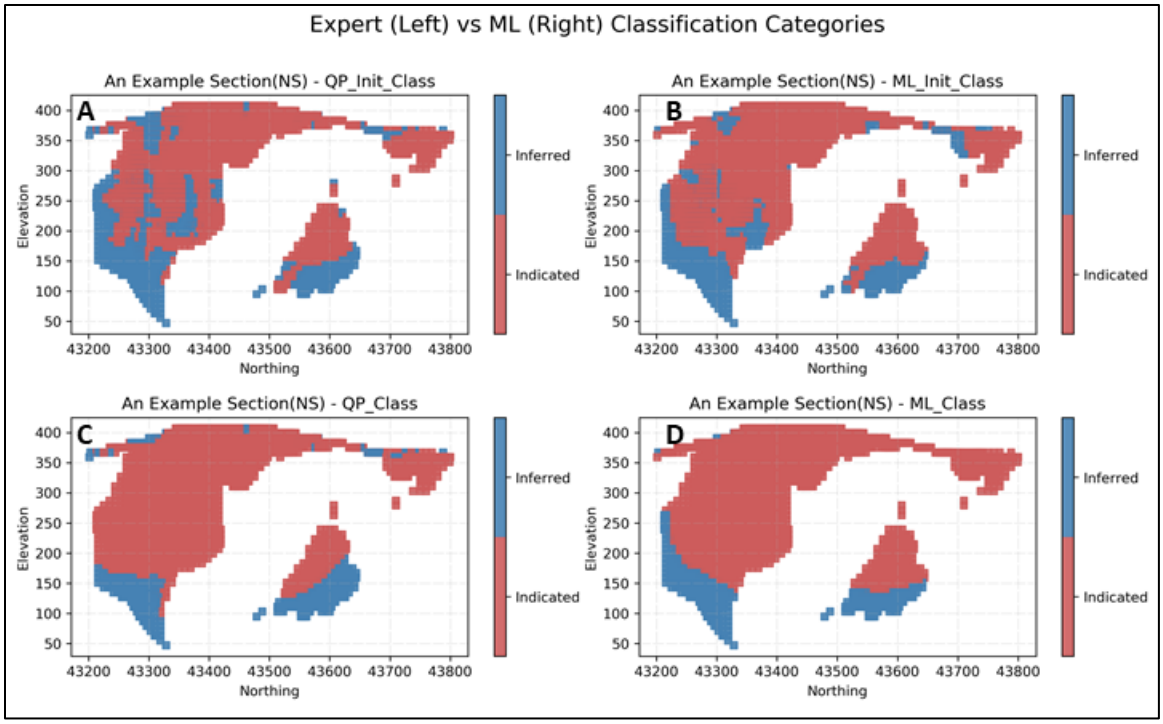


Figure 3.22 Comparison of QP (A and C) and ML (B and D) produced classification categories. A-B shows the initial classes and C-D shows the classes after smoothing applied.

It is shown that the proposed approach provides clusters that are consistent with the data, i.e. there is considerably good separation between clusters in terms of estimation parameters that are considered as proxies to the confidence level.

The maximum and minimum number of drill holes and composites used in estimation are constrained by the kriging plan which is set by the practitioner. Therefore, there may be some artificial limits for these parameters. For example, if a block is estimated with a total of 20 samples and the other was estimated with a total of 10 samples, one may think that the data density in the vicinity of the first block is higher. However, it should be noted that the maximum number of samples might be lower in the initial estimation passes due to the kriging plan to limit the smoothing in the estimates and can be relaxed in subsequent estimation pass passes. Therefore, the number of drillholes and composites may not be a direct indication of the data density. In addition

to this, the data density itself cannot be directly translated to confidence as long as the usage of the data is limited by the kriging plan. In any case, very low numbers in these two parameters can be considered as a proxy to low confidence in the estimation, which is represented by the inferred classes of all cases.

SVM with RBF kernel is a useful tool to conduct the smoothing operation as it allows the practitioner to control the amount of smoothing through its two hyper-parameters. The trade-off between consistency and the smooth boundaries can be adjusted by using accuracy and percent change metrics presented here.

These results show that machine learning approach produces results that are comparable to a classification done by a Qualified Person. It is transparent in that it can be reproducible given parameters and provides time efficiency. Once the parameters are set, total run time for the classification and smoothing algorithms are around 6 hours (3 hours per each). In addition, it improves the consistency between classes since thresholds to separate classes are not explicitly defined and inferred from the multivariate distribution of the estimation parameters.

3.3.6. Conclusion

We have presented a machine learning approach to resource classification task which normally requires several decisions to be made by a qualified person. Traditionally, the practitioner defines the thresholds of estimation parameters to draw the boundaries between classes. Qualitative and quantitative comparison of the results show that the automatic approach presented in this study can be an alternative tool to use in resource classification. Results are comparable to a qualified person classification and do not require to determine thresholds explicitly to draw the boundary between classes.

There are still important decisions that must be made by the practitioner, e.g. the total number of classes, and the highest level of confidence class, such as measured or indicated. Once these decisions are made, the proposed approach can produce results that are consistent with the data and in an automatic manner, i.e. with the minimum amount of human interaction. Classification results are reproducible and auditable.

Acknowledgements

Dr. Ortiz acknowledges the support of the Natural Sciences and Engineering Council of Canada (NSERC), funding reference number RGPIN-2017-04200 and RGPAS-2017-507956. Cevik would like to thank the Ministry of National Education Turkey for providing the scholarship funding of the M.A.Sc. The authors acknowledge Mitacs Accelerate and SRK Consulting Inc. (Canada) for supporting this research project.

Chapter 4 Conclusion and Future Work

4.1. Discussion of results

Researches of machine learning applications in the mineral resource sector have emerged especially in the last decade. This coincides with the contemporary challenges of the mining industry, such as declining grades, increasing depth of deposits, combined with the ever-increasing demand for the raw materials. Therefore, utilizing all available data efficiently, which is a well-known property of machine learning algorithms in many fields, becomes more relevant. Yet, machine learning applications have not been fully adopted by the industry. The main scope of the research presented in this contribution is to explore and develop machine learning workflows that are adoptable by the industry.

Two case studies are presented in Chapter 3, first one being related to applying knowledge discovery tools to lithogeochemical data and the second being automation of the mineral resource classification through a combination of unsupervised and supervised learning tools. A summary of the main conclusions of the two case studies and a brief discussion of the effects of data democratization and open source applications are presented to conclude this section.

4.1.1. Knowledge discovery through unsupervised exploratory tools

The workflow presented in the first case study in Chapter 3 aims to utilize unsupervised learning methods as exploratory data analysis tools. The main conclusions can be drawn as follows:

- Geochemical data are compositional in nature and this should be addressed in statistical analysis. For example, Aitchison distance, a distance specific to compositional data, is used to impute missing values with a k-Nearest Neighbor approach, where Euclidian distance is a common first-choice for many applications in non-compositional data. Similarly,

associations between variables, e.g. elements or oxides and observations are determined by principal component analysis instead of conventional correlation analysis after a centered-log transformation, a transformation specific to alleviate the effects of closed nature of the compositional data.

- Both linear (e.g. PCA) and non-linear (e.g. t-SNE) dimension reduction / projection methods are useful to reveal interesting groupings in the data. Non-linear methods can enhance the groupings between the observations; therefore, they can be used to focus the attention to the geologist to interesting parts of the data and formulate additional relevant research questions accordingly. However, they are not interpretable in terms of how they combined the variables in their projection axes. PCA, on the other hand, provides means of interpretation of the underlying processes that generated those groupings in the data both in observations (drill core samples), and variables (elements). Therefore, using linear projections in conjunction with the non-linear ones can help to gain insights in exploratory studies.
- The knowledge discovery cannot be realized without an understanding of the problem domain, i.e. domain expertise is essential to make use of all the potential benefits that an unsupervised exploratory analysis could provide. The key step in the Vazante study was the exploration and interpretation of the principal components that relate to the geological processes which potentially associated with the mineral deposits in the Vazante District. Here, association of the elements in PC1 was interpreted as reflecting a pre-enrichment process (ground preparation); and associations in PC2 and PC6 were interpreted as a signal for a depletion process in potential sources for ore-related elements in a sedimentary-hosted base metal belt These interpretations call for an understanding of the combination of

mineralogy, petrology and mineral systems (source, transport and deposition of metals) that form the mineral deposits.

4.1.2. Automation of resource classification

A workflow that was presented as the second case study (Chapter 3) aims to assist practitioners in the resource classification task by automating the process, while still leaving important decisions to the expert. The main conclusions can be summarized as follows:

- Resource classification is a subjective process, and the automation process presented in this study does not aim to remove the subjectivity, rather it aims to improve the consistency in the results given the data. In fact, the opinion of an expert in classification task cannot be removed as long as there are sources of uncertainties that are not captured by numerical models: uncertainty in geological interpretations, uncertainty in analysis methods or sample types, uncertainties due to sampling handling procedures adopted in the project site. Expert opinion facilitates the decision making where there is no data (or model) to quantitatively assess the specific sources of uncertainties,
- Qualitative data can be integrated to the proposed workflow as the MLA used in calculating block similarities can handle both numerical and categorical data.
- Once the important decisions are made, such as the highest confidence class in resources, the classification process can be automated with minimum human interaction. Results are comparable to expert classification.
- Block-by-block classification yields classes that are not spatially contiguous. This is usually caused by the use of multiple kriging passes during estimation, which generates multiple sets of distributions for the output variables (estimation variance, number of samples / drill holes used in estimation, etc.). This leads to distributions of these variables

that are not completely separable. These similarities between different estimation passes increase the inter-class similarities and result in spotty resource classes which are not desired for downstream applications in mining. Hence the spatial continuity needs to be imposed during clustering or as a post-process. In both cases, there is a trade-off between spatial continuity and consistency of the confidence classes given the input variables. The proposed approach provides the expert with means to control this trade-off between spatial continuity and class consistency.

4.1.3. Data-democratization and open-source applications

As mentioned previously, there is an ever-increasing research interest on machine learning applications in almost every field, including the mineral resource sector. This is partly sourced from the recent trends of transforming the information and expertise to a more accessible format for most of the individuals in the world through internet. This is reflected in papers that are freely accessible or free massive open online courses (MOOC) prepared by various universities, as well as open-source applications, software and libraries. In particular, programming languages such as R and Python, and freely available scientific libraries of these, have had a large impact in the popularization of machine learning methods.

Specific to the mining industry, some companies or geological surveys have organized crowdsourcing competitions which stimulated developments of many tools to solve mining related problems, e.g. *Integra Gold Rush Challenge (Integra Gold Corp, 2016), Explorer Challenge (Oz Minerals, 2019), The Gawler Challenge (The Government of South Australia, 2020).*

All aspects of this research are benefited from the open source information and applications. The workflows presented here are generated in open-source programming languages, namely R and

Python, and by using many different specialized libraries such as *robCompositions (R), NumPy, pandas, matplotlib* and *scikit-learn (Python),* among others. Moreover, most of the applications reviewed within the scope of this thesis were conducted using similar open source tools. This suggests that democratization of the information and contributing to open-source applications are essential to increase the development of novel applications and tools that might solve the contemporary challenges of the mining industry. From a geometallurgical point of view, the modeling tools reviewed and presented here demonstrate a significant potential to allow global optimization of the entire mining value chain. This can be done by linking all the stages, automating the processes and decisions, and improving our understanding of the interconnections between processes.

4.2. Main contributions of the thesis

This section briefly discusses the main contributions of the thesis:

- A high-level overview of some of the most recent applications of MLA in the literature, related to mineral exploration and resource sector, is provided. There are a wealth of case studies and applications in the literature. Focus was given to the applications that present interesting and novel uses of MLA and relevant to the problems addressed in this thesis. A generic workflow for good practice for MLA applications in spatial data is provided as an outcome of this review.
- An original knowledge discovery workflow for mineral exploration studies is presented in the first case study in Chapter 3, which led to a paper submitted to a peer-reviewed journal. The interpretation of the multivariate analysis revealed patterns that can be related to a wide-spread sub economic enrichment in ore-related elements in the underlying siliciclastic units of Vazante District (Serra do Garrote Formation, SGF). These patterns

indicated that SGF has favorable geochemical signatures to be considered as a potential source rock for the known deposits located in the upper carbonate units. If further studies support the hypothesis that SGF was indeed the source of the metals, patterns that are revealed in this study can be used as exploration criteria in the belt.

- An original workflow for mineral resource classification with state-of-art use of RF in unsupervised mode, with potentially a significant impact on the mining industry is presented. This led to another paper currently in preparation. This demonstrates that processes usually driven by humans can be automated through ML applications. This reduces the time and work involved and improves time efficiency and consistency.

4.3. Future Works

Some challenges were identified in this study but were not addressed either because it was out of scope or because an alternative approach was preferred:

- By default, commonly used supervised machine learning applications (reviewed in Chapter 2) do not consider spatial correlation between data points (autocorrelation) and are trained on each training data individually. Neglecting spatial autocorrelation yields suboptimal models for spatial predictions especially if the data is scarce, e.g. many applications of prospectivity mapping where most of the evidential layers, i.e. co-variables, have a resolution of a hundred meters (e.g. Torppa et al., 2019; Carranza et al., 2015). There are attempts to integrate spatial-contextual information as auxiliary variables (see Cracknell, 2014 and references therein, Hengl et al., 2018), but they are not widely adopted. Therefore, studies related to integration of the spatial context to machine learning applications are an exciting research topic to improve predictive performance of the MLA in spatial data.

- In the first case study (knowledge discovery), it is shown that non-linear methods can be useful to support the identification of interesting groupings in the data. However, the nonlinear projection method used (t-SNE) was chosen solely based on its demonstrated success in literature on applications related to other fields. Alternatives were not investigated thoroughly, or no sensitivity has been conducted related to its hyper-parameters. In order to achieve optimum results for knowledge discovery by using non-linear transformations, further studies to address alternative approaches and sensitivities to hyper-parameters are required.
- In the second case study (automated resource classification), spatial continuity is provided with a post-process smoothing application. Adding spatial constraints to clustering may eliminate the post processing step, and reduces the amount of the human intervention, possibly at the cost of less control of the expert on the trade-off between spatial continuity and class consistency in terms of confidence variables.
- Validation techniques, both for unsupervised and supervised applications are an open subject for spatial machine learning algorithms. For example, possible issues for supervised learning applications in generating validation data sets related to spatial autocorrelation have been recently recognized by researchers (Roberts et al., 2017; Pohjankukka et al., 2017). Accounting for the spatial correlation in these validation sets leads to spatially distinct groups, unlike the approach taken in the past where validation sets were being prepared by random (shuffled) sampling. This resulted in overly optimistic performance measures. There are few studies to evaluate the performance of unsupervised learning methods in the spatial context where the ground truth is not available. Developing robust

performance metrics that can assimilate different data types, and different objectives would impact the development of useful unsupervised MLA.

References

Agterberg, F. P., Bonham-Carter, G. F., Cheng, Q. M., & Wright, D. F. (1993). Weights of evidence modeling and weighted logistic regression for mineral potential mapping. Computers in Geology, *25*, 13-32.

Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), 44(2), 139-160.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Log ratio analysis and compositional distance. Mathematical Geology, 32(3), 271-275.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Arndt, N. T., Fontboté, L., Hedenquist, J. W., Kesler, S. E., Thompson, J. F., & Wood, D. G. (2017). Future global mineral resources. Geochemical Perspectives, 6(1), 1-171.

Battalgazy, N., & Madani, N. (2019). Categorization of mineral resources based on different geostatistical simulation algorithms: A case study from an iron ore deposit. Natural Resources Research, 28(4), 1329-1351.

Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. Science, 363(6433).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L. (2003). Manual--Setting Up, Using, And Understanding Random Forests V4.0. Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf

Carranza, E. J. M. (2008). Geochemical anomaly and mineral prospectivity mapping in GIS. Elsevier.

Carranza, E. J. M. (2011). Analysis and mapping of geochemical anomalies using log ratio-transformed stream sediment data with censored values. Journal of Geochemical Exploration, 110(2), 167-185.

Carranza, E. J. M., & Laborte, A. G. (2015). Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Computers & Geosciences, 74, 60-70.

Carvalho, I. A., Olivo, G. R., Moura, M. A., & Oliveira, G. D. (2017). Fluid evolution in the southern part of the Proterozoic Vazante Group, Brazil: Implications for exploration of sedimentary-hosted base metal deposits. Ore Geology Reviews, 91, 588-611.

Caté, A., Perozzi, L., Gloaguen, E., & Blouin, M. (2017). Machine learning as a tool for geologists. The Leading Edge, 36(3), 215-219.

Chayes, F. (1960). On correlation between variables of constant sum. Journal of Geophysical research, 65(12), 4185-4193.

CIM (2019). Estimation of Mineral Resources & Mineral Reserves Best Practice Guidelines. Canadian Institute of Mining. Retrieved from https://mrmr.cim.org/en/best-practices/estimation-of-mineral-resources-mineral-reserves.

Coombes, J., Fahey, G., & Stoker, P. T. (2014) Overview – Classification and reporting. Mineral Resource and Ore Reserve Estimation—The AusIMM Guide to Good Practice. Australasian Institute of Mining and Metallurgy, second edition, Monograph 30, 767-770.

Cordeiro, P. F., Oliveira, C. G., Paniago, L. N., Romagna, G., & Santos, R. V. (2018). The carbonate-hosted MVT Morro Agudo Zn-Pb deposit, central Brazil. Ore Geology Reviews, 101, 437-452.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Cracknell, M. J., & Reading, A. M. (2013). The upside of uncertainty: Identification of lithology contact zones from airborne geophysics and satellite data using random forests and support vector machines. Geophysics, 78(3), WB113-WB126.

Cracknell, M. J., & Reading, A. M. (2014). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Computers & Geosciences, 63, 22-33.

Cracknell, M. J. (2014). Machine learning for geological mapping: Algorithms and applications (Doctoral dissertation, University of Tasmania).

Dardenne, M. A. (2000, August). The Brasília fold belt. In Tectonic Evolution of South America. 31st International Geological Congress, Rio de Janeiro (Vol. 231, p. 263).

Deutsch, C. V., Leuangthong, O., & Ortiz, J. M. (2007). Case for geometric criteria in resources and reserves classification. Transactions-Society For Mining Metallurgy And Exploration Incorporated, 322, 1.

Dohm, C. (2005). Quantifiable mineral resource classification: A logical approach. In: Leuangthong O., Deutsch C.V. (eds) Geostatistics Banff 2004. Quantitative Geology and Geostatistics, vol 14, 333-342. Springer, Dordrecht

Douce, A. E. P. (2016). Metallic mineral resources in the twenty-first century. I. Historical extraction trends and expected demand. Natural Resources Research, 25(1), 71-90.

Emery, X., & Lantuéjoul, C. (2006). Tbsim: A computer program for conditional simulation of three-dimensional gaussian random fields via the turning bands method. Computers & Geosciences, 32(10), 1615-1628.

Emery, X., Ortiz, J. M., & Rodríguez, J. J. (2006). Quantifying uncertainty in mineral resources by use of classification schemes and conditional simulations. Mathematical Geology, 38(4), 445-464.

Fernandes, N. A., Olivo, G. R., & Layton-Matthews, D. (2019a). Siliciclastic-hosted zinc mineralization in the Proterozoic Vazante–Paracatu District, Brazil: Implications for metallogeny and sources of metals in sediment-hosted base metal systems. Ore Geology Reviews, 114, 103139.

Fernandes, N. A., Olivo, G. R., Layton-Matthews, D., Voinot, A., Chipley, D., & Diniz-Oliveira, G. (2019b). Geochemistry and provenance of siliciclastic rocks from the Mesoproterozoic Upper

Vazante Sequence, Brazil: Insights on the evolution of the southwestern margin of the São Francisco Craton and the Columbia Supercontinent. Precambrian Research, 335, 105483.

Fouedjio, F., Hill, E. J., & Laukamp, C. (2018). Geostatistical clustering as an aid for ore body domaining: case study at the Rocklea Dome channel iron ore deposit, Western Australia. Applied Earth Science, 127(1), 15-29.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Goodfellow, W. D., & Lydon, J. W. (2007). Sedimentary exhalative (SEDEX) deposits. Mineral deposits of Canada: A synthesis of major deposit types, district metallogeny, the evolution of geological provinces, and exploration methods: Geological Association of Canada, Mineral Deposits Division, Special Publication, 5, 163-183.

Goovaerts, P. (1999). Impact of the simulation algorithm, magnitude of ergodic fluctuations and number of realizations on the spaces of uncertainty of flow properties. Stochastic Environmental Research and Risk Assessment, 13(3), 161-182.

Granek, J. (2016). Application of machine learning algorithms to mineral prospectivity mapping (Doctoral dissertation, University of British Columbia).

Groves, D. I., & Santosh, M. (2015). Province-scale commonalities of some world-class gold deposits: implications for mineral exploration. Geoscience Frontiers, 6(3), 389-399.

Grunsky, E. C. (2010). The interpretation of geochemical survey data. Geochemistry: Exploration, Environment, Analysis, 10(1), 27-74.

Harris, J. R., & Grunsky, E. C. (2015). Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Computers & Geosciences, 80, 9-25.

Hedenquist, J. W., Arribas, A. & Gonzalez-Urien, E. (2000). Exploration for epithermal gold deposits. Reviews in Economic Geology, 13(2), 45-77.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, *6*, e5518.

Hood, S. B., Cracknell, M. J., & Gazley, M. F. (2018). Linking protolith rocks to altered equivalents by combining unsupervised and supervised machine learning. Journal of Geochemical Exploration, 186, 270-280.

Hood, S. B., Cracknell, M. J., Gazley, M. F., & Reading, A. M. (2019). Improved supervised classification of bedrock in areas of transported overburden: applying domain expertise at Kerkasha, Eritrea. Applied Computing and Geosciences, 3, 100001.

Hronsky, J. M. (2009). The exploration search space concept: key to a successful exploration strategy. Centre for Exploration Targeting Quarterly News, 8, 14-15.

Hron, K., Templ, M., & Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. Computational Statistics & Data Analysis, 54(12), 3095-3107.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, pp. 3-7). New York: Springer.

JORC Code (2012) Australasian code for reporting of exploration results, mineral resources and ore reserves. AusIMM, Melbourne, p 44

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8), 1544-1554.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. Statistical Data Analysis based on the L1 Norm. Y. Dodge, Ed, 405-416.

Kirkwood, C., Cave, M., Beamish, D., Grebby, S., & Ferreira, A. (2016). A machine learning approach to geochemical mapping. Journal of Geochemical Exploration, 167, 49-61.

Kuhn, S., Cracknell, M. J., & Reading, A. M. (2019). Lithological mapping in the Central African Copper Belt using Random Forests and clustering: Strategies for optimised results. Ore Geology Reviews, 112, 103015.

Large, R. R., Bull, S. W., McGoldrick, P. J., & Walters, S. G. (2005). Stratiform and strata-bound Zn-Pb-Ag deposits in Proterozoic sedimentary basins, northern Australia. Economic Geology, 100, 931-963.

Leach, D. L., Sangster, D. F., Kelley, K. D., Large, R. R., Garven, G., Allen, C. R., ... & Walters, S. (2005). Sediment-hosted lead-zinc deposits: A global perspective. Economic Geology, 100(3), 561-607.

Le Bas, M. J., Le Maitre, R. W., & Woolley, A. R. (1992). The construction of the total alkali-silica chemical classification of volcanic rocks. Mineralogy and Petrology, 46(1), 1-22.

Le Maitre, R. W., & Ferguson, A. K. (1978). The CLAIR data system. Computers & Geosciences, 4(1), 65-76.

Le Maitre, R. W. (1984). A proposal by the IUGS Subcommission on the Systematics of Igneous Rocks for a chemical classification of volcanic rocks based on the total alkali silica (TAS) diagram: (on behalf of the IUGS Subcommission on the Systematics of Igneous Rocks). Australian Journal of Earth Sciences, 31(2), 243-255.

Li, S., Yang, C., Sun, H., & Zhang, H. (2019). Seismic fault detection using an encoder–decoder convolutional neural network with a small training set. Journal of Geophysics and Engineering, 16(1), 175-189.

Liaw, A. and Wiener, M. (2002). Classification and Regression by Random Forest. R News 2(3), 18-22.

Lishchuk, V., Koch, P. H., Ghorbani, Y., & Butcher, A. R. (2020). Towards integrated geometallurgical approach: Critical review of current practices and future trends. Minerals Engineering, 145, 106072.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.8.

McCuaig, T. C., & Hronsky, J. M. (2014). The mineral system concept: the key to exploration targeting. Society of Economic Geologists Special Publication, 18, 153-175.

McGladrey, A. J. (2014). The Integration of Physical Rock Properties, Mineralogy and Geochemistry for the Exploration of Large Hypogene Zinc Silicate Deposits: A Case Study Of The Vazante Zinc Deposits, Minas Gerais, Brazil (Doctoral dissertation, Queen's University).

McGladrey, A. J., Olivo, G. R., Silva, A. M., Oliveira, G. D., Neto, B. B., & Perrouty, S. (2017). The integration of physical rock properties, mineralogy and geochemistry for the exploration of large zinc silicate deposits: A case study of the Vazante zinc deposits, Minas Gerais, Brazil. Journal of Applied Geophysics, 136, 400-416.

Misi, A., Azmy, K., Kaufman, A. J., Oliveira, T. F., Sanches, A. L., & Oliveira, G. D. (2014). Review of the geological and geochronological framework of the Vazante sequence, Minas Gerais, Brazil: Implications to metallogenic and phosphogenic models. Ore Geology Reviews, 63, 76-90.

Mitchel, T. M., & Learning, M. (1997). McGraw-Hill. New York.

Monteiro, L. V. S., Bettencourt, J. S., Spiro, B., Graca, R., & de Oliveira, T. F. (1999). The Vazante zinc mine, Minas Gerais, Brazil; constraints in willemitic mineralization and fluid evolution. Exploration and Mining Geology, 8(1-2), 21-42.

Monteiro, L. V. S., Bettencourt, J. S., Juliani, C., & de Oliveira, T. F. (2006). Geology, petrography, and mineral chemistry of the Vazante non-sulfide and Ambrósia and Fagundes sulfide-rich carbonate-hosted Zn–(Pb) deposits, Minas Gerais, Brazil. Ore Geology Reviews, 28(2), 201-234.

Monteiro, L. V. S., Bettencourt, J. S., Juliani, C., & de Oliveira, T. F. (2007). Nonsulfide and sulfide-rich zinc mineralizations in the Vazante, Ambrósia and Fagundes deposits, Minas Gerais, Brazil: mass balance and stable isotope characteristics of the hydrothermal alterations. Gondwana Research, 11(3), 362-381.

Neves, L. P. (2011). Características descritivas e genéticas do depósito de Zn-Pb Morro Agudo, Grupo Vazante (Doctoral dissertation, Dissertação de mestrado, Universidade de Brasília).

Olivo, G., Monteiro, L., Baia, F., Slezak, P., Carvalho, I., Fernandes, N., Oliveira, G.D., Botura Neto, B., McGladrey, A., Silva, A.M., Moura, M.A., Layton-Matthews, D., & Moura, M. (2018). The Proterozoic Vazante hypogene zinc silicate district, Minas Gerais, Brazil: a review of the ore system applied to mineral exploration. Minerals, 8(1), 22.

Ortiz JM (2019) Geometallurgical modeling framework, Predictive Geometallurgy and Geostatistics Lab, Queen's University, Annual Report 2019, paper 2019-01, 6-16.

Ortiz JM, Cevik SI, Avalos S, Kracht W, Leuangthong O (2020) Machine learning and deep learning in predictive geometallurgical modeling, PDAC, Toronto, ON, March 4, 2020

Ortiz, J. M., & Deutsch, C. V. (2003) A Practical Way to Summarize Uncertainty for Classification, Centre for Computational Geostatistics, Report Five, University of Alberta, Sept. 2003, 14 p.

Palacio-Niño, J. O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. arXiv preprint arXiv:1905.05667.

Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). Modeling and analysis of compositional data. John Wiley & Sons.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. —on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 60(359-367), 489-498.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of Machine Learning Research, 12, 2825-2830.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. International Journal of Geographical Information Science, 31(10), 2001-2019.

Reimann, C., Filzmoser, P., & Garrett, R. G. (2002). Factor analysis applied to regional geochemical data: problems and possibilities. Applied Geochemistry, *17*(3), 185-206.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroder, B., Thuiller, W., et al. (2017). "Crossvalidation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." Ecography, 40(8): 913–929.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804-818.

Romary, T., Rivoirard, J., Deraisme, J., Quinones, C., & Freulon, X. (2012). Domaining by clustering multivariate geostatistical data. In Geostatistics Oslo 2012 (pp. 455-466). Springer, Dordrecht.

Romary, T., Ors, F., Rivoirard, J., & Deraisme, J. (2015). Unsupervised classification of multivariate geostatistical data: Two algorithms. Computers & Geosciences, 85, 96-103.

Rossi, M. E., & Deutsch, C. V. (2013). Mineral resource estimation. Springer Science & Business Media.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500), 2323-2326.

SAMREC. 2016. South African Mineral Resource Committee. The South African Code for the Reporting of Exploration Results, Mineral Resources and Mineral Reserves (the SAMREC Code. 2016 Edition. https://www.samcode.co.za/samcode-ssc/samrec

Schubert, E., & Rousseeuw, P. J. (2019). Faster k-Medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In International Conference on Similarity Search and Applications (pp. 171-187). Springer, Cham.

Shi, T., & Horvath, S. (2006). Unsupervised learning with random forest predictors. Journal of Computational and Graphical Statistics, 15(1), 118-138.

Sillitoe, R. H. (2010). Porphyry copper systems. Economic Geology, 105(1), 3-41.

Silva, D. S. F., & Boisvert, J. B. (2014). Mineral resource classification: a comparison of new and existing techniques. Journal of the Southern African Institute of Mining and Metallurgy, 114(3), 265-273.

Slezak, P. R., Olivo, G. R., Oliveira, G. D., & Dardenne, M. A. (2014). Geology, mineralogy, and geochemistry of the Vazante Northern Extension zinc silicate deposit, Minas Gerais, Brazil. Ore Geology Reviews, 56, 234-257.

Stephenson, P. R., Allman, A., Carville, D. P., Stoker, P. T., Mokos, P., Tyrrell, J., & Burrows, T. (2014). Mineral resource classification – It's time to shoot the 'spotted dog'! Mineral Resource and Ore Reserve Estimation—The AusIMM Guide to Good Practice. Australasian Institute of Mining and Metallurgy, second edition, Monograph 30, 799-804.

Stephenson, P. R., & Stoker, P. T. (2001). Classification of mineral resources and ore reserves. Mineral Resource and Ore Reserve Estimation—The AusIMM Guide to Good Practice. Australasian Institute of Mining and Metallurgy, Monograph 23, 653-660.

Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z., & Hu, Z. (2020). Data-Driven Predictive Modelling of Mineral Prospectivity Using Machine Learning and Deep Learning Methods: A Case Study from Southern Jiangxi Province, China. Minerals, 10(2), 102.

Talebi, H., Mueller, U., Tolosana-Delgado, R., Grunsky, E. C., McKinley, J. M., & de Caritat, P. (2019). Surficial and deep earth material prediction from geochemical compositions. Natural Resources Research, 28(3), 869-891.

Templ, M., Hron, K., Filzmoser, P. (2011). robCompositions: an R-package for robust statistical analysis of compositional data. In V. Pawlowsky-Glahn and A. Buccianti, editors, Compositional Data Analysis. Theory and Applications, pp. 341-355, John Wiley & Sons, Chichester (UK).

Torppa, J., Nykänen, V., & Molnár, F. (2019). Unsupervised clustering and empirical fuzzy memberships for mineral prospectivity modelling. Ore Geology Reviews, 107, 58-71.

Van den Boogaart, K. G., & Tolosana-Delgado, R. (2013). Analyzing compositional data with R (Vol. 122). Heidelberg: Springer.

Wawruch, T. M., & Betzhold, J. F. (2005). Mineral resource classification through conditional simulation. In: Leuangthong O., Deutsch C.V. (eds) Geostatistics Banff 2004. Quantitative Geology and Geostatistics, vol 14, 479-489. Springer, Dordrecht.

Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1-23.

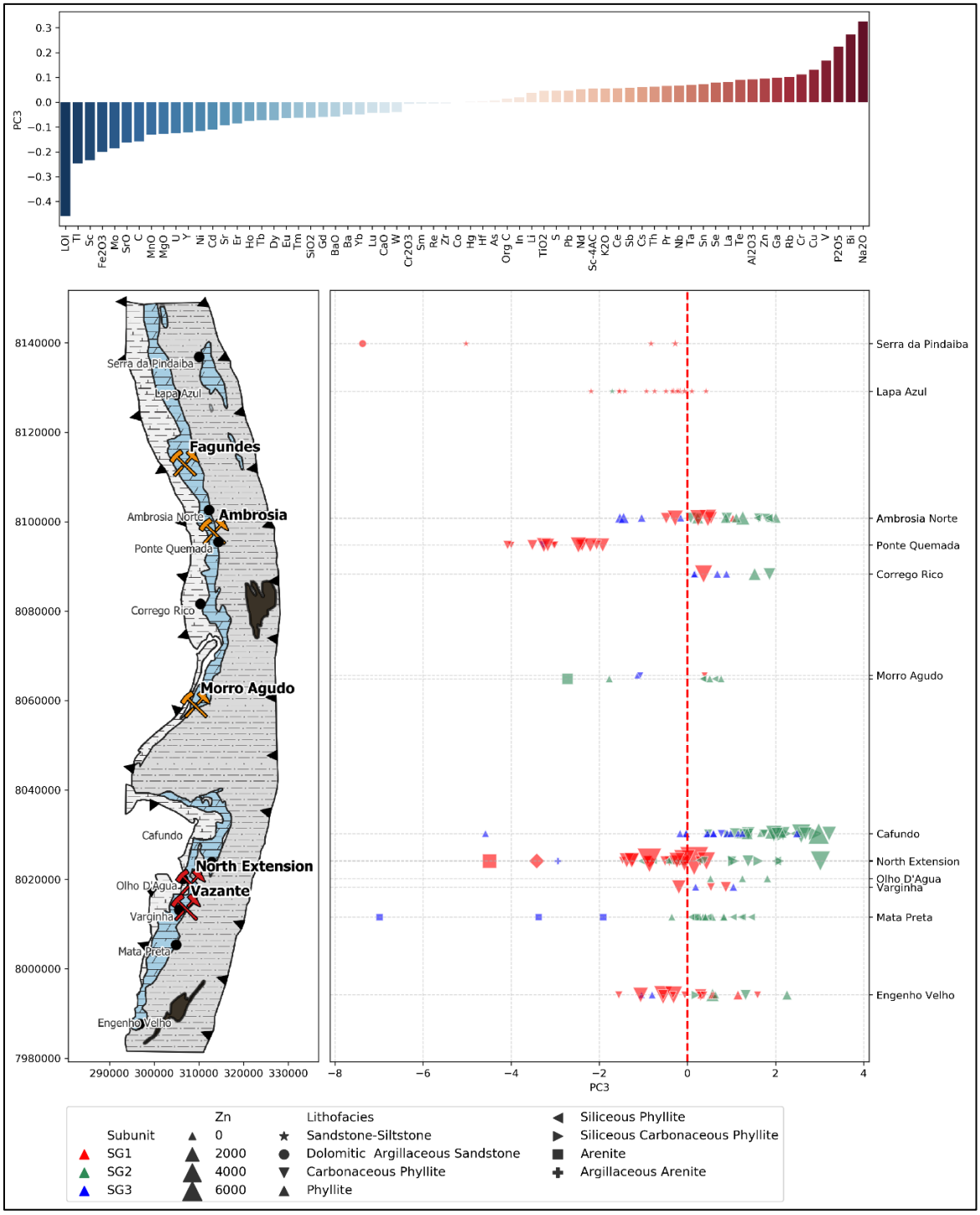
Wyborn, L. A. I., Heinrich, C. A., & Jaques, A. L. (1994, August). Australian Proterozoic mineral systems: essential ingredients and mappable criteria. In The AusIMM Annual Conference (Vol. 1994, pp. 109-115). AusIMM Darwin.

Yeomans, C. M., Shail, R. K., Grebby, S., Nykänen, V., Middleton, M., & Lusty, P. A. (2020). A machine learning approach to tungsten prospectivity modelling using knowledge-driven feature extraction and model confidence. Geoscience Frontiers.

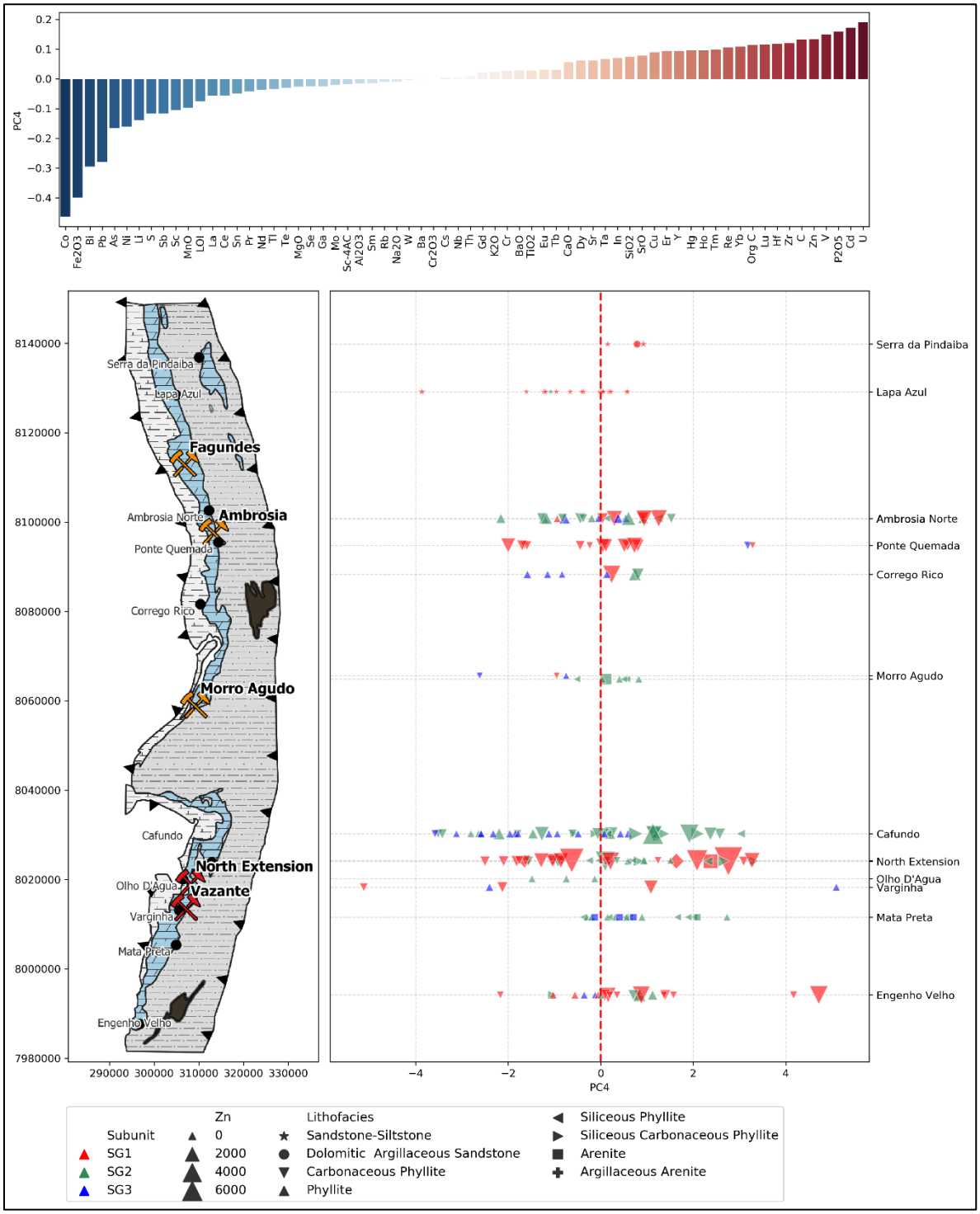
Zhang, Q., Yusifov, A., Joy, C., Shi, Y., & Wu, X. (2019). FaultNet: A deep CNN model for 3D automated fault picking. In SEG Technical Program Expanded Abstracts 2019 (pp. 2413-2417). Society of Exploration Geophysicists.

Appendices

A. Map shows the PC3 distribution of the samples in spatial context together with element loadings in the PC3, and the relative sample zinc contents. The geological map on the left is modified from Fernandes et al., 2019b.



B. Map shows the PC4 distribution of the samples in spatial context together with element loadings in the PC4, and the relative sample zinc contents. The geological map on the left is modified from Fernandes et al., 2019b.



C. Map shows the PC5 distribution of the samples in spatial context together with element loadings in the PC5, and the relative sample zinc contents. The geological map on the left is modified from Fernandes et al., 2019b.