UNIVERSITY OF TASMANIA

DOCTORAL THESIS

Machine Learning for Mineral Exploration: Prediction and Quantified Uncertainty at Multiple Exploration Stages

---

Author:
Stephen KUHN

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

ARC Hub for Transforming the Mining Value Chain
CODES - Centre of Ore Deposit and Earth Sciences
School of Natural Sciences

May 2021

Declaration of Authorship

I, Stephen KUHN, declare that this thesis titled, "Machine Learning for Mineral Exploration: Prediction and Quantified Uncertainty at Multiple Exploration Stages" and the work presented in it are my own.

Declaration of Originality. This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Authority of Access. The non-published content of the thesis (see below) may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

Statement Regarding Published Work Contained in Thesis. Chapters 3, 4 and 5 of this thesis are published under a Creative Commons Attribution (CC BY) licence. You are free to copy, communicate and adapt the work, so long as you attribute the author.

Statement of Co-Authorship
The following people and institutions contributed to the publication of work undertaken as part of this thesis:

Stephen Kuhn, Centre for Ore Deposit and Earth Sciences (CODES), School of Natural Sciences, University of Tasmania = **Candidate**

Matthew James Cracknell, Centre for Ore Deposit and Earth Sciences (CODES), School of Natural Sciences, University of Tasmania = **Author 1**

Anya Marie Reading, School of Natural Sciences (Physics), School of Earth Sciences, University of Tasmania = **Author 2**

Stephanie Sykora, First Quantum Minerals Ltd. = **Author 3**

Author Details and Their Roles

Paper 1, 'Lithological mapping using Random Forests applied to geophysical and remote sensing data: a demonstration study from the Eastern Goldfields of Australia', located in Chapter 3.

Paper 2, 'Identification of indicator lithologies in volcanic terrains in British Columbia by machine learning using Random Forests: the value of using a soft classifier', located in Chapter 4.

Paper 3, 'Lithological Mapping in the Central African Copper Belt using Random Forests and Clustering: Strategies for Optimised Results', located in Chapter 5.

The candidate was the primary author of each paper. Authors 1 and 2 contributed to the development, refinement and presentation of each paper. Author 3 contributed the foundational geological understanding of the geographical location in which the study comprising Paper 2 was conducted.

Signed: 30 Sept 2020

Date:

Anya M. Reading
Supervisor
School of Natural Sciences (Physics)
University of Tasmania

Signed:

Date: 30 Sept 2020

Sebastien Meffre
Head of Discipline, Earth Sciences
Centre for Ore Deposit and Earth Sciences
School of Natural Sciences
University of Tasmania

UNIVERSITY OF TASMANIA

Abstract

CODES - Centre of Ore Deposit and Earth Sciences
ARC Hub for Transforming the Mining Value Chain

Doctor of Philosophy

Machine Learning for Mineral Exploration: Prediction and Quantified Uncertainty at Multiple Exploration Stages

by Stephen KUHN

Machine learning describes an array of computational and nested statistical methods whereby a computer can 'learn' and subsequently make predictions or identify patterns in data. With the increasing volume and variety of numerical data in the geosciences, and widespread availability of the needed computing power, machine learning techniques are a logical addition to the numerous possible approaches that can be applied to the search for ore deposits.

The three core research chapters in this thesis develop the application of machine learning in the context of mineral exploration. Emphasis is placed on the Random Forests algorithm for mapping lithology in a range of settings and at a variety of stages in the exploration process. Information entropy is used to assist both in assessing and communicating any complex combinations, and potential inaccuracy, of classification results. Through the thesis, methods are employed with future practical usage in mind, such that machine learning may be used by the geologist (as domain expert) in an objective manner.

The first of these core studies uses the Random Forests algorithm to re-classify the solid geology lithology map of the Heron South project, located in the Eastern Goldfields of Western Australia. This study uses geophysical and remote sensing data, in the absence of geochemical samples and geological ground truthing with most of the project under transported cover. This is characteristic of an early stage, reconnaissance exploration project. A sparse training sample of 1.6 percent of the total area, is taken as training data, allowing much of the areas geology the freedom to be reclassified. This study demonstrates that Random Forests, with proper consideration given to sampling and training data selection, can be used effectively to produce or improve geological mapping in little-explored areas. Information entropy is shown to be valuable in predicting where classification was likely to be inaccurate or a region highly complex.

The second core study uses Random Forests to produce a solid geology map of the Kliyul porphyry prospect of British Columbia, Canada, using a fusion of available geophysical and geochemical data, typical of a greenfields stage exploration project. Soil and rock chip sample sites were taken as training data, used to classify the remainder of the project area. Assessment of the probability distributions produced using the Random Forests algorithm enabled regions with an elevated probability

of intrusions (a key indicator lithology) to be mapped, even where not observed in training data. The results of this study highlight the value of a soft, ensemble classifier such as Random Forests, and the value to be gained from an assessment of the spatial distribution of class probabilities as opposed to viewing a final map as a solution in isolation.

In the third and final core study, a range of training data sampling paradigms are tested in a data rich area located in the Domes region of the Central African Copper belt hosting the Sentinel (Ni) and Enterprise (Cu) deposits. This study simulates early and advanced stage exploration project maturity in incorporating a priori geological Information. It culminates in the use of Random Forests to undertake an objective audit of the present company geological map. Further to this, unsupervised clustering is used in the production of a geological map in the absence of training or constraint through identifying the natural grouping of data. The results of these studies highlight the importance of proper sample balancing and explore the repercussions of limited and/or non-representative training data. The use of the information entropy proxy is developed to identify where a classification may depart from the domain represented by training data. The ranking of input data that is performed in association with the Random Forests classification can be used to improve clustering results through optimising dataset selection.

Through the three core research chapters, a set of practical considerations and recommendations for explorers are provided. It is demonstrated that Random Forests can provide an objective audit and subsequent refinement of a pre-existing geological map. The expression of uncertainty using information entropy, and the assessment of class probabilities, can be used to appraise the results from the machine learning analyses. This includes validation in the case of complex outcome combinations, and generation of new insights. Ranking of input datasets via Random Forests can enhance understanding of data and improve both Random Forests classification results and improve clustering. With the proper selection of appropriate datasets, clustering (for example immobile trace elements) and scaling can indeed produce results that correspond well with lithology. Studies presented in this thesis use data from current/active exploration projects and methods are distilled to streamlined workflows using industry standard software and data formats. In summary, these methods, previously the domain of computer and data scientists, are now developed to be more widely accessible to mineral explorers.

Acknowledgements

This research was supported by the ARC Research Hub for Transforming the Mining Value Chain (project number IH130200004). I would acknowledge Andrew Foley (Gold Fields Australasia Ltd.) and Chris Wijns, Tim Ireland and Mike Christie (First Quantum Minerals Ltd.) for allowing use of their proprietary geoscience data.

It's not often I find the time I should to stop and reflect on the journey so far; so I would like to take this opportunity to thank all those along the way who helped get me to this point:

Thanks to my parents who always selflessly made the effort to give myself and my brother the opportunity and encouragement to pursue whatever path we chose.

Thanks to all my undergraduate lecturers at CODES, particularly: Andrew Tunks, Peter McGoldrick, Tony Crawford, Michael Roach and Mark Duffett. They likely do not know this, but their competence and enthusiasm inspired a teenager who would rather have been outside kicking a football than sitting in a classroom, to take up a career in exploration.

Thank you to all my old colleagues at Gold Fields for your friendship and mentoring, particularly Juan Barrera, Stanley Zutah, Mark Falconer, Dale Sanderson and Philip Brown. Special thanks Janet Tunjic and Justin Osborne, for giving me my first shot in the industry and to Andrew Foley for giving me the opportunity to work all over the world and pushing me to develop all aspects of my geophysical skill set: technical, scientific and importantly vocational/fieldcraft.

To Leo and the crew at the Weightlifting Academy of Tasmania. Beyond the enjoyment of the sport; the sophistication of programming and training methodology coupled with the hard work, character, discipline and consistency required for high performance (or as close as a 32 year old who squats too much and can't snatch to save his life can get…) set an example for me that I'll always try to live up to. Likewise, to Steve, Wayne, and Erik and everyone at Southern ITF.

Thanks to the Compute Earth group and the students and staff at CODES with whom I collaborated, for your time and company. Special thanks to my office mate Esi both for being good company and the opportunity to regularly dip my toes back into potential fields, my first love in geophysics.

To my PhD supervisors: Anya Reading and Matt Cracknell; where do I even start? I have not made this easy for you, especially after family circumstances forced my abrupt and somewhat premature move back to industry. Quite simply, without your support, patience, and guidance I would not be here now to write this at the front of a completed thesis.

Lastly, but certainly not least: Jyldyz, my wife. We have made it through just about every challenge a couple could ever have to face. You manage to put up with my cynical outlook and sardonic responses to most things… and remain the most cheerful and open person I have ever met. Your love and support through not just my candidature but the last 10 years has been nothing short of amazing. I could not have done it without you.

Contents

List of Figures

List of Tables

For my Father

Chapter 1

Introduction

1.1 Introduction

Data-driven methods are used across many aspects of human endeavour and present a new set of possible research approaches for the applied geosciences. The availability of multiple datasets and the ubiquity of powerful portable computing capability may be leveraged together as a transformational opportunity for the resources industries. Machine learning describes the process where a computer can learn from a, sometimes large, volume of high dimensional data and make meaningful inferences or predictions, with minimal human input (Mitchel, 1997; Dutton & Conroy, 1996). The production of an improved geological map, from airborne geophysical data, remote sensing data, and sparse ground observations, by machine learning techniques has been demonstrated (e.g. Cracknell, Reading & McNeill, 2014). The success of early studies, particularly using Random Forests (RF) has been developed in ongoing and related research (e.g. Cracknell & Reading, 2014; Cracknell, 2014; Harris & Grunsky, 2015; Yu et al., 2012).

A wide variety of data are routinely used in mineral exploration which, as in the demonstration study noted above, can be grouped into geophysical (often airborne) data, remote sensing (satellite) data, and geological/geochemical observations. These are available at differing scales and resolutions defined by the prior history of investigation, prospectivity, exploration stage and specific needs of a given project. This thesis will focus on lithological mapping of the Earth's surface, focussing on the priorities most likely for mineral exploration. In this context, the availability of data increases with the progression of a project from precompetitive airborne potential field geophysical and satellite data to "boots on ground" geological mapping, geochemical sampling and high-resolution geophysical surveys. With each addition in terms of data variety, data resolution and the increasing availability of direct geological observations, the ability to produce more robust lithological maps through conventional means increases. As such, usage of machine learning must demonstrate an addition of value for the given exploration stage. For example, the production of a 'first pass' geological map from sparse observations is a very useful result in a preliminary, remote desktop study for a given project. Where the geology is well mapped and understood, however, a product of this nature does not add any significant value. In this case of a subsequent exploration stage, the ability to refine existing mapping while revealing, quantitatively, the associated uncertainty could be significant. Thus, there could be numerous ways that machine learning might add value to a mineral exploration project.

1.1.1 Research proposition and aims

The research described in this thesis aims to extend the application of machine learning to a variety of geological settings, hosting a range of mineral commodities, and at different stages of exploration project maturity. Three new machine learning case studies are presented that use industry-sourced data and interpretations developed in discussion with the industry geologist or geochemist (Figure 1.1). Taken together, the case studies enable a more thorough appraisal of the effectiveness of machine learning for lithology prediction and information evaluation in the context of mineral exploration. Frequently encountered challenges for exploration geologists such as data sparsity, data non-uniformity, outcrop absence and sampling biases are addressed.

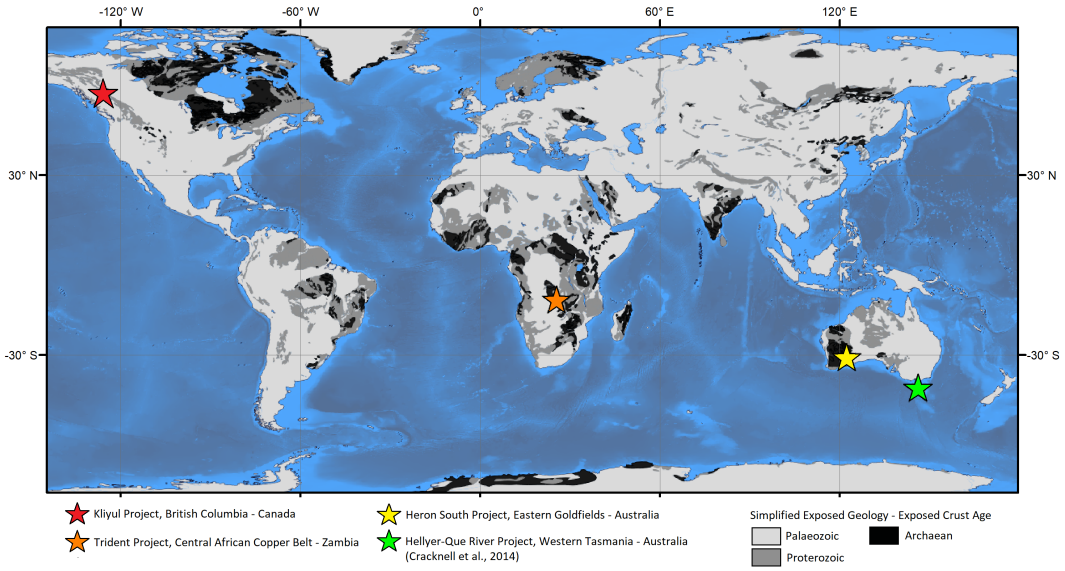


FIGURE 1.1: Simplified world geology map, grouped by age of exposed crust, over seafloor bathymetry. The three studies comprising this thesis along with work by Cracknell et al. (2014) are shown.

The overarching research proposition under investigation is that machine learning, the supervised RF and unsupervised clustering algorithms, if used appropriately, can produce, or refine a lithological map in varied geological settings and for a variety of mineral exploration contexts.

Specifically, through this research, I aim to:

1. Expand the range of demonstration studies for lithological map production or refinement using RF classification.

2. Progress prediction evaluation and other metrics for knowledge generation. This includes the use of quantified uncertainty and probabilistic assessment of classification output.

3. Identify, at various stages of project maturity, where lithological mapping aided by machine learning can add value in mineral exploration.

In performing this research, I focus on pragmatic workflows and methodology which are understandable and accessible to a range of geologists and geoscientists working in exploration. This is reflected in the usage of industry standard software packages for data handling, compilation, visualisation, and map production. Additionally, this research, while preserving academic rigour, honours the realistic conditions, practicalities, and outcome expectations of the mineral exploration industry over the life cycle of an exploration project. Accurate lithological mapping is a key component of mineral exploration and underpins all interpretation of numerical datasets. As such, this research addresses the task of prediction of lithology, or refinement of prior lithological mapping, in the context of the mineral explorer looking to define geology accurately to identify target locations for further investigation. It progresses the ability to improve prior mapping to prioritise or minimise further "boots on ground" mapping. These objectives, specific to each case study, will be progressed through the three core research chapters.

1.1.2 Thesis structure

To address the aims stated above, this thesis comprises a literature review chapter followed by three core science chapters, each of which constitutes a manuscript published in an international peer-reviewed journal. Each of these core chapters represents a study, or series of studies, that applies machine learning for lithological map production or refinement in a specific context with regards to geological setting, data availability and exploration project maturity (Figure 1.1). This thesis concludes with a synthesis and discussion of key outcomes, with respect to the research aims as outlined in this chapter. The contents of each of the following chapters are as follows:

Chapter 2: Background. A review of machine learning, focusing on geological mapping applications. The specific algorithms used in this thesis: RF, k-means and Self-Organising Maps will be described as will the concept of information entropy, the chosen uncertainty metric (proxy) for classification outputs in this thesis. The second section of this chapter provides a review of the ore deposit types which constitute exploration targets in the study areas comprising this thesis.

Chapter 3: Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia, published in GEOPHYSICS, 2017. This study focuses on the outcomes of a RF classification using geophysical and remote sensing data, simulating an early stage exploration setting. Additionally, this study investigates the relationship between uncertainty, in the form of information entropy (Shannon, 1948) and classification inaccuracy.

Chapter 4: Identification of intrusive lithologies in volcanic terrains using Random Forests soft classification: a demonstration study from British Columbia, published in GEOPHYSICS, 2020. This study investigates the application of RF for lithological map production in a porphyry setting. This study outlines the value of assessing RF internal probability statistics in conjunction with the final classification result in order to detect subtle intrusive bodies: potential hosts for mineralisation in conjunction with an early stage geological mapping and sampling programme.

Chapter 5: Lithological mapping in the Central African Copper Belt using Random Forests and clustering: Strategies for optimised results, published in ORE GEOLOGY REVIEWS, 2019. This research comprises a series of studies located in the Central African Copper Belt. These studies progressively simulate a variety of exploration project stages, identifying how machine learning, in the form of supervised classification or clustering, can be used to add value at each stage to produce, improve or audit lithological mapping.

Chapter 6: Synthesis. A synthesis and discussion of the outcomes of Chapters 3, 4 and 5 presented with respect to the research aims presented in this thesis. A summary of practical considerations for explorers using machine learning, and strategies to handle them, as identified in the research comprising this thesis will be provided.

Lastly, this chapter will include some discussion on the direction of future work

integrating an increasing role for machine learning into the toolbox of mineral explorers and geological mapping.

Chapter 7: Conclusions. A final statement of the findings of the research described in this thesis.

As the published chapters were written to stand alone as individual scholarly works, there is necessarily some repetition of background material throughout the thesis.

References

Cracknell, M. (2014). *Machine Learning for Geological Mapping: Algorithms and Applications. ARC Centre of Exellence in Ore Deposits (CODES) School of Physical Sciences (Earth Sciences)*, University of Tasmania.

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

Cracknell, M., Reading, A. and McNeill, A. (2014). Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests^{tm} and Self-Organising Maps, *Australian Journal of Earth Sciences* **61**: 287–304.

Dutton, D. and Conroy, G. (1996). A review of machine learning, *The Knowledge Engineering Review* **12**(4): 341–367.

Harris, J. and Grunsky, E. (2015). Predictive lithological mapping of Canada's north using Random Forest classification applied to geophysical and geochemical data, *Computers and Geosciences* **80**: 9–25.

Mitchell, T. (1997). *Machine Learning*, McGraw-Hill, New York.

Yu, L., Porwal, A., Holden, E. and Dentith, M. (2012). Towards automatic lithological classification from remote sensing data using support vector machines, *Computers and Geosciences* **45**(0): 229–239.

Chapter 2

Background

This chapter presents background material, which forms the foundation for the case studies in subsequent chapters. Firstly, a review of previous work concerning machine learning in the geosciences is provided. This is followed by sections that outline the algorithms used in one or more of the case studies. Secondly, the diversity of tectonic settings for ore deposits are outlined with a view to the choice of exploration target locations for these case studies. Finally, a short summary of priorities for lithology mapping for mineral exploration is provided. Thus, the context for the applied research presented in this thesis is given.

Machine learning algorithms (MLAs) provide an objective data-driven approach to lithological mapping and minerals targeting. The research detailed in this thesis utilises Random Forests (RF), and to a lesser extent, clustering in the form of k-means and Self-Organising Maps (SOM). The research applies this approach to a variety of ore deposit styles in order to predict lithology from disparate geophysical, geochemical and remote sensing data. This will require an evaluation of type of data, data sampling strategy and the stage of a project in which the approach will deliver the most beneficial return. To understand the usefulness of this approach, a background, both in the underlying techniques and the ore deposit models to which it will be applied, is required.

In the first section of this chapter, I will define the concept of MLAs and outline the geological problem they are being deployed to solve. RF supervised classification, k-means and SOM unsupervised clustering algorithms are then reviewed. This review is complemented by a summary of the implementation of these algorithms for geological mapping applications, including the recent, seminal work in which the combined RF/SOM approach to lithological mapping was developed. Also summarised is the role of uncertainty in geological applications, in particular the use of information entropy (H; Shannon, 1948).

An understanding of end-member ore deposit models to which these techniques will be applied is critical to understanding how the outputs of MLAs can generate meaningful geological knowledge. The general characteristics and setting of the deposit styles under investigation: Archaean granite-greenstone orogenic Au, porphyry (Cu-Au-Mo) and epithermal (Ag-Au) systems will be described. Key properties of these deposit styles and how they are identified in geophysical, geochemical and remote sensing data are examined. This will provide a framework of observed characteristics that underpins mapping and mineral targeting through the combined RF/SOM approach. A summary is presented, juxtaposing the quantifiable observation criteria for the suite of each deposit style. An understanding of these criteria will allow for the products of the combined

RF/SOM approach to be interpreted in the context of an ore deposit environment, turning data-driven predictions into geological knowledge and thus enhancing opportunities for mineral discovery.

2.1 Machine Learning Methods

Machine learning describes a process whereby a computer can learn from data, with minimal human input. In many situations this improves prediction performance or ability to generalise to other tasks (Mitchel, 1997; Dutton & Conroy, 1996). MLAs provide a data driven, often inductive means of solving problems, exploiting the ever-increasing performance of computers, by identifying patterns in high-dimensional space. The concept of machine learning is not new. Alan Turing, for example, well known in the public domain for contributions to breaking German cyphers during the second world war, posed the question "Can Machines think?" in a paper published in Mind 70 years ago (Turing, 1950), speculating on many aspects of the potential of computers to imitate thought and learning processes. The increasing performance and accessibility of computing resources have led to the development of a multitude MLAs.

MLAs can be grouped broadly into two forms: supervised and unsupervised. An MLA can be described as supervised if it utilises training data with known labels. The supervised MLA learns a set of rules describing the relationship between the training data and a priori class labels. In classification problems (including those that form the basis of our research), a MLA is then able to predict labels for new, previously unseen data. The general process of training a supervised classifier (Hastie et al., 2009; Kotsiantis, 2007) is shown in Figure 2.1. Conversely, an MLA can be described as unsupervised where prior information defining classes is not provided. Instead, an unsupervised algorithm searches for natural groups, clusters, patterns or responses in the data (Hastie et al., 2009; Kotsiantis, 2007).

2.1.1 Geoscience data for machine learning classification

MLAs have become an attractive option for classification (and regression) problems in geoscience (e.g. Bierlein, 2007; Cracknell & Reading, 2014, 2015; Fraser & Dickson, 2008; Harris & Grunsky, 2015; Yu et al., 2011). Geoscientific data are generally not considered "Big Data" in the truest sense, not meeting the criterion for sheer volume. For example, the first documented reference to "Big Data" (Cox and Ellsworth, 1997) defines the term as datasets being too large to fit in memory or even on hard disk of a system. Geoscientific data is regularly however, highly disparate, high-dimensional and highly spatially variable. Supervised classifiers allow geoscientists to define lithological classes, or other geological subdivisions, while retaining the ability to objectively learn a set of rules for making that classification, free of user bias (Kotsiantis, 2007) at this stage of analysis.

A variety of geochemical, geophysical and remote sensing data are required for both conventional or MLA based geological prediction. These data are collected by industry or government due to their utility in mineral exploration and used to map and target a variety of features relating to various mineralisation styles discussed later in this chapter. In this sense, pre-selection and acquisition of these data

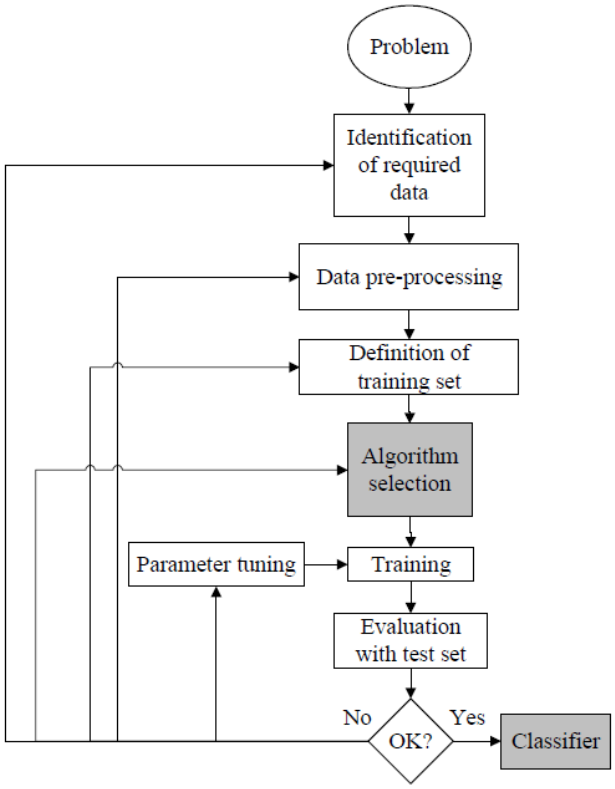


FIGURE 2.1: Generalised structure of the supervised machine learning problem (from Kotsiantis, 2007).

propagate a degree of subjectivity/bias into what is otherwise, and throughout this body of work, described as an objective, data driven process.

In order to produce a complete lithological map, data must be interpolated or imputed into locations at which no direct measurement was made. Some data types lend themselves well to such a process: ASTER (Abrams, 2000) Landsat (National Aeronautics and Space Administration, 2006) or Shuttle radar topography mission (SRTM; Farr et al, 2003) are collected at regular increments and are pervasive. aeromagnetic, radiometric and electromagnetic geophysical data are typically acquired at a rapid rate along a flight lines which are commonly separated by tens to hundreds of meters. Methods such as bi directional splines or minimum curvature are pervasively used to interpolate these data to a regular grid. gravity data, being commonly equi-spaced to line separation double station separation are also typically interpolated via minimum curvature (Briggs, 1974). It should be understood that a degree of smoothing is inherent in all datasets prior to interpolation due to the effects of vector quantisation when measuring a continuous value at discrete intervals. While each sample is attributed to a pixel of the same interval, the reading was taken at the central point and assumed to be representative of the pixel. An increase in sample resolution results in the ability to define higher spatial frequencies. This however is an attribute of data as collected and delivered by industry or the government and not a parameter that can be modified for the purposes of the study. Smoothing may be required where noise is excessive, usually in the form of a low pass filter. There are several alternate choices to a strict low pass filter which mitigate ringing (Gibbs phenomenon) such as cosine roll-off, upward continuation of potential field data, Butterworth low-pass,

or a simple convolution such as Hanning. These options are all industry standard, viable, public domain techniques. Although somewhat subjective, this is currently considered industry best practice. The threshold for low pass filtering is also subjective and varies between datasets; determined by the noise apparent in the data, the noise specified by the equipment manufacturer (often calibrated for the particular instrument) and qualitative assessment of the solid earth of which the data are representative. Any error produced or rather incompletely mitigated at this stage will cascade through to classification uncertainty.

Conventionally, data representing geochemistry/litho-geochemistry have been used to identify lithotype, alteration zones and other anomalous regions relating to a mineralising event. While these remain powerful tools, they are limited by the practitioners' ability to define a series of meaningful relationships between data in a 2D or 3D framework. Unsupervised clustering algorithms may have the potential to aid practitioners by discovering alteration signatures or other ore deposit characteristics in high-dimensional space. These expressions are highly variable and difficult to define with the precision required for supervised classification or conventional mapping techniques (Bierlein et al., 2007). Studies (as will be discussed in the following section) that have used various machine learning approaches consistently conclude that MLA outperform more conventional statistical or manually weighted classification methods.

The research comprising this thesis will focus on the deployment of RF for lithological mapping in a range of mineral exploration contexts and geological settings. Clustering will also be investigated in some instances as an alternative to supervised classification in map production. The following sections of this chapter will introduce the algorithms used in this thesis: RF, k-means and SOM.

2.1.2 Random Forests

RF is an ensemble classification algorithm developed by Breiman (2001). The classifier constructs multiple randomised decision trees or a 'forest'. The class assigned to a given sample or instance within the data is defined by a majority vote cast by all the trees in the forest (Breiman, 2001). This allows for superior classification performance when compared to decision trees and other single classifiers, as these are more prone to bias and over-fitting (Hastie et al., 2009).

Randomness is introduced into the algorithm at two stages. Firstly, a subset comprising a pre-defined number of input variables, selected at random from all available input variables, is used to split instances at each node of a decision tree. Secondly, bagging (Breiman, 1996) is used to modulate the training data available to each decision tree in the forest. Bagging, also called bootstrap aggregation, obtains training data for each decision tree by randomly sampling, with replacement, a subset of the training dataset with the number of samples equal to the training dataset. Due to the replacement of samples, some are selected in multiple instances while others may not be sampled. Testing by Breiman (1996) showed that an average of $\tilde{6}3\%$ of instances are included in the training subset, while the remaining or "out-of-bag" samples ($\tilde{3}7\%$) are used to internally evaluate trained classification models. Each node within the decision trees is split using a threshold that improves the homogeneity of the child node. The Gini index, as described by Breiman et al. (1984) is used to provide a measure of information

purity of the child node relative to the parent node. Thus, the Gini Index defines the threshold that produces the maximum reduction in class heterogeneity from parent to child node. The process is effectivly non-linear in that variable assessed, and value at which it is split is considered independently of that which preceded it (except in partition of the total dataset that has passed through). As such, no assumption relating elements comprising a multi-element sample summing to 100% is made and thus the problem of closure (Aitchison, 1982) is mitigated in the way RF is deployed in this work. The need to consider closure when analysing the statistical relationship between elements in a given sample, in machine learning applications is described in detail by Hood et al. (2019).

RF calculates an estimate of class membership probability (CMP) that describes the probability each sample belongs to candidate classes in the training data. The RF CMP is defined as the quotient of the frequency of votes for a class divided by the number of trees in the forest (Hastie, et al., 2009). This information is subsequently used to estimate the uncertainty associated with the classification of a given dataset and will be discussed further in the following section (2.1.2). Breiman (2001), determined that the two parameters which determined the (out-of-bag) error rate within a forest are: the correlations between any two trees; and the strength of the individual trees comprising the forest. Stronger trees, i.e. those with a lower error rate contribute to an overall lower error rate for the forest. In contrast, increasing correlation between trees increases the overall error rate of the forest. Breiman (2001) went on to show that reducing the number of input variables reduces both the strength of trees and their correlations. There is an optimal range in the number of variables selected for a given training dataset, balancing the two error sources results in a low error rate for the forest. RF is largely immune to direct over-fitting and multicolinearity and thus a stable minimum error can be reached even with the addition of a high number of variables and/or noisy data. The inclusion of such data can however, contribute to a reduction in the strength of individual trees and increase the chance of correlation between trees. Reducing the number of input variables also contributes to ease of interpretation and improved computational efficiency.

It is necessary to remove highly correlated input variables and identify a minimum number of relevant input variables without negatively impacting accuracy. Correlation between variables can be assessed via standard statistical correlation metrics such as Pearson's or Spearman's correlation coefficient and removed when above a defined threshold. RF provides an internal measure of variable importance by taking an average of the decrease in the Gini Index for all nodes in a forest split using that variable (Breiman 2001; Cutler et al. 2007; Cracknell et al., 2014). The objective is to include only as many variables as is necessary until a point of diminishing returns is reached and the error rate stabilises. This can be achieved through manual experimentation or is popularly achieved through a recursive variable elimination approach (e.g. Kuhn et al. 2012; Pedregosa et al., 2011).

The number of trees in a forest has an effect on overall accuracy. As with the number of input variables, this reaches a point of diminishing returns with regard to a given study and training dataset (e.g. Cracknell et al., 2014; Harris & Grunsky, 2015; Rodriguez-Galiano et al., 2014; Waske, 2009). In all these examples, it was observed that a certain number of decision trees are required to reach a stable error minima, beyond this point, additional trees add to the computational cost of the

algorithm without improving accuracy and are therefore superfluous.

RF has been shown to achieve similar accuracy to other classification algorithms while having the advantage of being easier to use (e.g. Cracknell & Reading, 2013; Hastie et al., 2009). This makes RF a good candidate for use by geoscience practitioners as computational and mathematical literacy are not a limiting factor in wider adoption of the method.

RF has been increasingly applied to the problem of lithological classification. Waske et al. (2009) compared RF and another popular MLA, Support Vector Machines (SVM; Vapnik, 1998), to standard classifiers in the context of mapping lithology using hyperspectral imagery. They concluded that both methods achieved significantly more accurate results than standard classifiers. While in that instance, SVMs marginally outperformed RF, it was noted by the authors that RF remained an attractive option due to high accuracy and relative ease of use. Cracknell & Reading (2013) assessed RF and SVM for lithology mapping and the identification of lithological contacts and zones of structural complexity. They discovered that RF, in addition to an excellent overall performance, produced more usable outputs. High uncertainty (discussed in the next section) was associated with incorrectly classified samples, giving a robust measure of confidence in the accuracy of classifications. Additionally, unlike SVM results, areas of high uncertainty calculated from RF were spatially proximal to geological boundaries and zones of high structural complexity, which in turn could be a valuable tool in targeting exploration efforts. A further, rigorous study by Cracknell and Reading (2014) compared RF with other MLAs as applied to a lithological mapping problem. Their results showed that RF marginally outperformed other MLAs such as SVMs. Despite small differences in accuracy RF was able to produce these results with simpler input parameters, less sensitive to tuning and at less computational cost. The authors also noted that with the use of increasingly spatially dispersed training data, the performance gap between RF and other MLAs increased.

Cracknell and Reading (2014), deployed RF in conjunction with SOM to great effect in Tasmania's economically significant Mount Read Volcanics. The authors were able to accurately map the lithology of the area using a small training dataset (less than 2% of the instances from the available geological map) and by optimising input variables though a variable importance ranking process. RF was also able to identify unmapped features in the area, potentially of exploration significance. Harris & Grunsky (2015) utilised a similar approach, deploying RF for lithological mapping in northern Canada. They tested two training data selection scenarios: one based on lake sediment geochemical sample locations and another based on field mapping observation stations. Both approaches produced meaningful results, with gamma ray spectrometry and geochemical data (abundance of 60 elements) produced the best predictions. They conclude that RF is of value as a first pass mapping tool or in future focusing effort into areas where there is a mismatch between predicted geology and legacy maps.

2.1.3 k-means

k-means (Lloyd, 1957) is perhaps the simplest clustering algorithm; both conceptually and operationally, partitioning data on the basis of similarity as

defined by proximity to the nearest of a defined number of group means (Macqueen, 1967). A given number (k) of centroids are seeded in the data space and the distance between each data point and each centroid calculated. Each data point is assigned to the group defined by the mean to which it is nearest, in doing so, dividing the dataspace via Voronoi partitioning. Centroids are re-defined as the mean of all data points assigned to that group. All data points which now lie closer to an alternate centroid are re-allocated to that cluster. This process is re-iterated until stable clusters are achieved. Where applied in this thesis, silhouette analysis (Rousseeuw, 1987) is used to define the mathematically optimal number of clusters. Silhouette provides a measure of the dissimilarity of points within each cluster, as compared to dissimilarity to the nearest neighbouring cluster. Random seeding of starting centroids can result in a large number of iterations required to reach a satisfactory result, convergence in a local error minima or an inability to converge at all where the number of iterations is capped. Where deployed in this thesis, the k-means++ (Arthur & Vassilvitskii, 2007) variant of the k-means algorithm is used. k-means++ controls seeding of starting centroids, avoiding coincident centroids or those seeded as outliers relative to the range expressed in data. This produces superior processing performance and accuracy.

2.1.4 Self-Organising Maps

SOM are a class of unsupervised MLA proposed and formalised by Kohonen (1998, 2002). This author defines a process by which higher dimensional observations are mapped onto a 2D manifold or "constrained topological map" (Hastie, et al., 2009). This process, as outlined below and described in subsequent literature, includes applications relating to spatially distributed geoscientific datasets (e.g. Bedini, 2009; Bierlein et al., 2008; Cracknell et al., 2014; Cracknell & Reading, 2014; Cracknell, Reading & de Caritat, 2015; Fraser & Dickson, 2007; Klose, 2006). The following description draws primarily from Bierlein et al. (2008) and the sources cited therein.

The initial step in SOM, requires the dimensions (rows and columns) of the desired map to be defined and hence the total number of seed-nodes. A competitive step is employed whereby a process of vector quantisation and subsequent measures of vector similarity, such as Euclidean distance, are used to allocate an input sample to the best matching seed node/vector within a given radius (Figure 2.2). The seed node and all other nodes within a defined radius are modified to more closely resemble a given input sample. This step is performed for each input sample.

The process is iterative, with input data shown to the seed-vectors, each time with a decreasing radius within which seed-vectors are modified and a smaller modification permitted. Trained seed-vectors represent the characteristics of associated input data. In the resulting 2D map, input data are represented by the nearest trained seed-nodes. The process of mapping from n dimensional space (n being the number of variables comprising the input data) to 2D space preserves the topology of the input data: samples that were close in nD space, remain close in 2D space. This preservation of topology facilitates some ability to conceptually perceive the structure and organisation within the nD dataset.

SOMs have been increasingly applied to a variety of geoscientific problems. For example: Bierlein et al. (2008) used SOM to successfully define groups corresponding to ore deposit styles within a multivariate database comprising

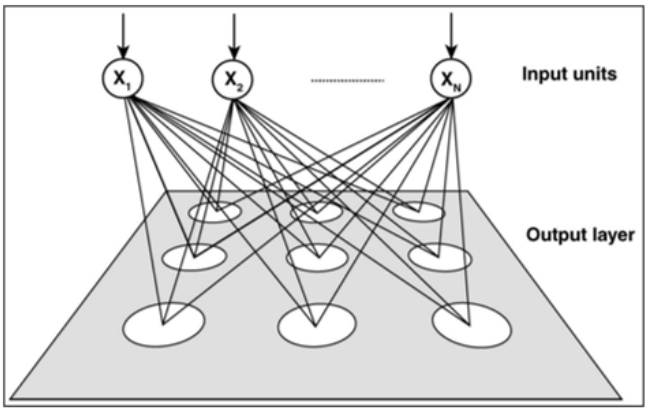


FIGURE 2.2: Representation of the structure of a SOM, mapping input variables (x_1–x_n) onto a 2D constrained topological map (Bierlein et al., 2008).

information on geochemistry, ore and alteration mineralogy, and ore geometry. These groups were linked with known deposit types and allowed the definition the subtleties and variability associated with host rock or location within each deposit style. These authors further examined the relationship between these groups and structure, though various statistical analyses. Tayebi & Tangestani (2015), used SOM to map the abundance of various alteration minerals in the Masashim volcano in Iran, achieving an accuracy of 83% when validated. Cracknell, Reading & de Cariat (2015) used SOM with a variety of geophysical and remote sensing datasets to classify the regolith over the Australian continental landmass. By combining this information with known mineral occurrences, they were able to produce continent-scale, mineral prospectivity maps for several commodities. Cracknell, Reading & McNeill (2014) combined SOM with RF to add value to the mapping and classification process via the identification of sub-units within volcanic classes predicted by RF.

2.1.5 The role of uncertainty in geological applications

Uncertainty is pervasive across geoscientific data and analysis (and indeed scientific inquiry in general) being introduced via a wide variety of sources. As such, a great number of anecdotal and quantitative definitions and frameworks have been used in its description and measurement. Mann (1993), building upon the work of Cox (1982) proposed a widely used framework defining three types of uncertainty. Type 1 is defined by error and bias in measurement. Type 2 comprises stochasticity and inherent randomness in a measured variable. Type 3 includes incomplete or incorrect knowledge. Wellmann et al. (2010) further adapted these definitions in a form well suited to a geospatial context. That being Type 1: the degree of bias or error in measurement, Type 2: inherent stochastic and randomness manifesting as uncertainty in interpolation (or imputation) of values between measured data points; and Type 3: conceptual uncertainty comprising incomplete, incorrect or imprecise understanding of the geological or structural context represented by data. Numerous strategies exist to combat Type 1 and Type 2 uncertainty as described above. Type 1 can be mitigated through robust QAQC of measurement apparatus and methods, measurement repetition to generate statistically significant and verifiable results and rigour in accurately documenting equipment error, resolution and detection limits. Type 2 uncertainty, as applied to

2D and 3D interpolation of geoscientific data, has been widely addressed in the field of geostatistics. Directional Analysis of semi-variance, domaining of data where possible, and variations on Krigging are widely considered provide a superior alternative to simple proximity weighted estimates (Chiles & Delfiner, 2012). These methods invariably produce a measure of uncertainty through calculation of variance/co-variance during Krigging at any given point (Chiles & Delfiner, 2012; Goovaerts, 1997). Type 3 uncertainty is difficult if not impossible to fully mitigate as this would require a complete and correct knowledge of the geology in question. The use of RF presents the opportunity to take a probabilistic approach to calculation of uncertainty associated with classification prediction. Such approach captures elements of both Type 1 and Type 2 uncertainty as they effectively cascade through to any final predictions based upon interpolated data. Prior research has demonstrated both qualitatively, through visualisation of results and quantitatively that the spatial distribution of high uncertainty relates to poor predictions (Kuhn et al., 2018) geological complexity (Kuhn et al., 2018; Cracknell & Reading, 2014; Wellmann & Regenauer-Leib, 2012) and lithological contacts (Cracknell & Reading, 2014). Cracknell & Reading (2014) further demonstrated the application of statistical rigour applied to non-ensemble based algorithms, permuting data selection, demonstrating a wide scope for use of probabilistic approach to uncertainty in MLA classification assessment. As such, a probabilistic method can be used to analyse Type 3 uncertainty manifesting as the validity of a RF classification or model. As such this generalised, comprehensive approach to describing uncertainty will be described in detail below and used throughout the studies comprising this thesis.

2.1.6 Uncertainty from RF

RF, in addition to a class label for each instance/pixel in a dataset, produces a class membership probability (CMP). This is in the form of a vector p_c comprising probabilities, with a length equal to the number of possible classes, which sum to 1. These outputs can be used to calculate a meaningful measure of the uncertainty or "fuzziness" of the classification for each point in a dataset. Two common metrics used to quantify the uncertainty associated with this distribution of class probabilities (Cracknell & Reading, 2013 and the sources reviewed therein; Goodchild et al., 1994; are H and variance. These methods are equally applicable to 2D or 3D geological modelling, given it is the class probabilities attributed to a data instance/pixel/voxel and not its geometry that are used in calculations of uncertainty. As such, 2D and 3D examples will be discussed interchangeably. As H has garnered attention in recent studies on the uncertainty associated with geological models (e.g. Wellmann & Ragenauer-Lieb, 2012), this method will be described. H was first defined by Shannon (1948) as:

$$H = -k \sum_{i=1}^{n} p_i log p_i$$

where p is a probability of each possible class: i, with n defining the total number of possible classes, k is an arbitrary constant to define unit of measure if required.

The objective when using H (or any uncertainty calculation) is to provide a system whereby a maximum value is achieved when all possible classes are equally probable in a cell and a minimum value when a single class has a 100% probability of being present and all other classes 0% (Figure 2.3). An important consideration

when using H (or any such uncertainty metric) on highly dimensional datasets is whether a normalisation factor should be applied. Cracknell and Reading (2013) for example, calculated variance, normalised to a scale of 0-1 whereas Leung, Goodchild & Lin (1993) also used an extra denominator, namely, a division by the number of classes present at a given sample, to normalise their results.

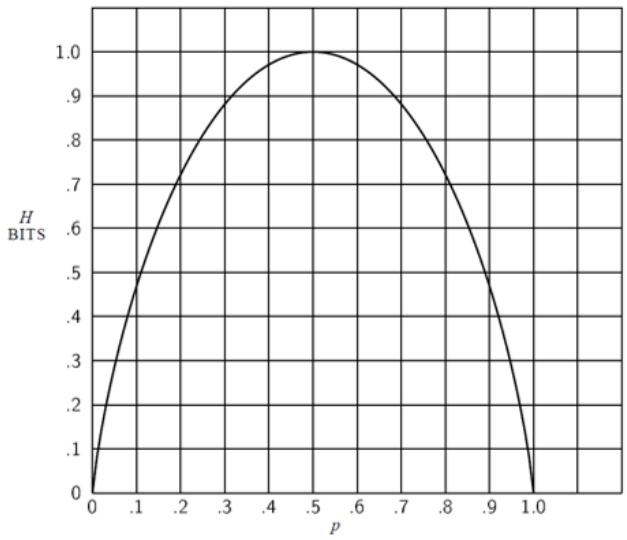


FIGURE 2.3: An illustration of the behaviour of H in a simple 1 bit (2 class) scenario (from Shannon, 1948). Note that H is zero when class 1 or 0. H increases towards a maxima when both classes exhibit an equal probability of occurrence.

Normalisation by the number of possible classes generates a consistent value, providing a measure describing each instance in the context of whether it is approaching its maximum or minimum uncertainty. This allows a classified dataset to be examined qualitatively for areas of high uncertainty in a relative sense. It does not however, accurately describe the absolute magnitude of uncertainty. Wellmann et al. (2012), advocated for a non-normalised approach where monotonicity is preserved, thus H is higher when a higher number of equally probable classes are present. It should be noted that the authors do also discuss the value of a normalised approach in some situations. A non-normalised approach, method is better able to distinguish cells of higher absolute complexity for example, five similarly probable classes as opposed to those where a class label was derived from, for example, two similarly probable likely classes. In a normalised (0-1) approach, both of these cases would approach 1. Conversely, instances with a low number of possible classes will present with relatively low uncertainties, even if, within the context of that instance, all classes are equally probable and therefore very difficult to accurately label. This can be a disadvantage as cells with a higher number of classes present may, as a function of the distribution of probabilities amongst those classes, may have been more reliably classified but retain high entropy. This in turn may be confused with cells with low H due to a low number of classes, albeit with a less reliable class label. These properties are discussed in the sources cited previously in this section and the choice of approach is largely determined by the objectives of each study and the aspect of probability distribution of classes the authors are trying to demonstrate. Both paradigms will be investigated further in the current study. A normalised approach, as per Cracknell et al. (2014) could be a better proxy of class label accuracy, while a non-normalised approach as per

Wellmann et al. (2012) may better describe potential absolute complexity at each point in a model. Thresholds will be explored in the context of a non-normalised approach, eliminating classes that are possible but exhibit a negligible probability. This provides an opportunity to extend on the previous research by providing a better means of maintaining monotonicity while eliminating redundant, low probability variables may inflate H relative to cells with a lower number of classes.

Irrespective of the choice of uncertainty measure, or the choice between/combination of a normalised versus non-normalised approach, measures of uncertainty are critical to the value of the machine learning approach to geological classification. Uncertainty provides a means of critically evaluating classified geological models produced by RF. Uncertainty can provide a proxy for incorrect classification (e.g. Kuhn et al., 2018; Cracknell et al., 2014), a measure of complexity (Wellmann et al., 2012) and as a vector to defining lithological boundaries. The use of uncertainty measures, in particular, H, will form an important component of the current research.

2.1.7 Combining Random Forests with clustering

Recent research (Cracknell, Reading & McNeill, 2014) demonstrated the effectiveness of combining RF with SOM. This approach combines the advantages of supervised classification and unsupervised clustering. RF was selected due to the strong performance in previous research (Cracknell and Reading, 2013). This work highlighted both the effectiveness of RF as a classification tool and the utility of uncertainty measurements calculated from its outputs. Using RF, Cracknell et al. (2014) were able to map a prospective volcanic hosted massive sulphide (VHMS) terrain into 21 lithological classes with an accuracy of 78.41.8% (based on exact 95% confidence limits). An assessment of input variables identified those datasets most important to mapping lithologies in the study area; valuable knowledge for future decision making in where to target often limited exploration expenditure. Measurements of uncertainty were useful to additional critical evaluation of the process. Areas of high uncertainty showed a strong correlation with misclassified samples and pointed towards zones of geological contacts or structural complexity. This ability to identify complex areas is a useful tool to guide exploration efforts in geologically complex areas and reduce unnecessary expenditure on areas that are easily defined through RF. SOM was subsequently deployed in order to differentiate natural groups within the data. Via this approach, they were able to identify a number of zones within the basaltic and andesitic units present in the region. These clusters were shown to be indicative of possible overprinting alteration, important to VHMS exploration and targeting. The study (Cracknell et al., 2014) demonstrated the viability of each method as well as the value in combining both the supervised classifier RF with the unsupervised SOM in a VHMS setting. Following from this success, it necessary to expand the use of this approach to other economically significant ore deposit types. The current study will assess if this approach can be equally effective in other settings and what modifications may be required to generate robust results.

2.2 Ore Deposit Models

Ore deposit models describe the settings and characterisation of mineralisation. All deposit models are defined by the common element of accumulated metals, representing and enrichment relative to a background signal. The degree to which individual ore systems conform to these models is highly variable. Nonetheless, ongoing research attempts to classify and describe ore systems using idealised models based primarily on their formation, timing, driving processes, favourable locations, tectonic setting and diagnostic characteristics as they appear in the modern rock record. Figure 2.4 for example, illustrates the general geological setting for several deposit styles included in this review. These generalised models provide a basis for mineral exploration and targeting exploration and targeting.

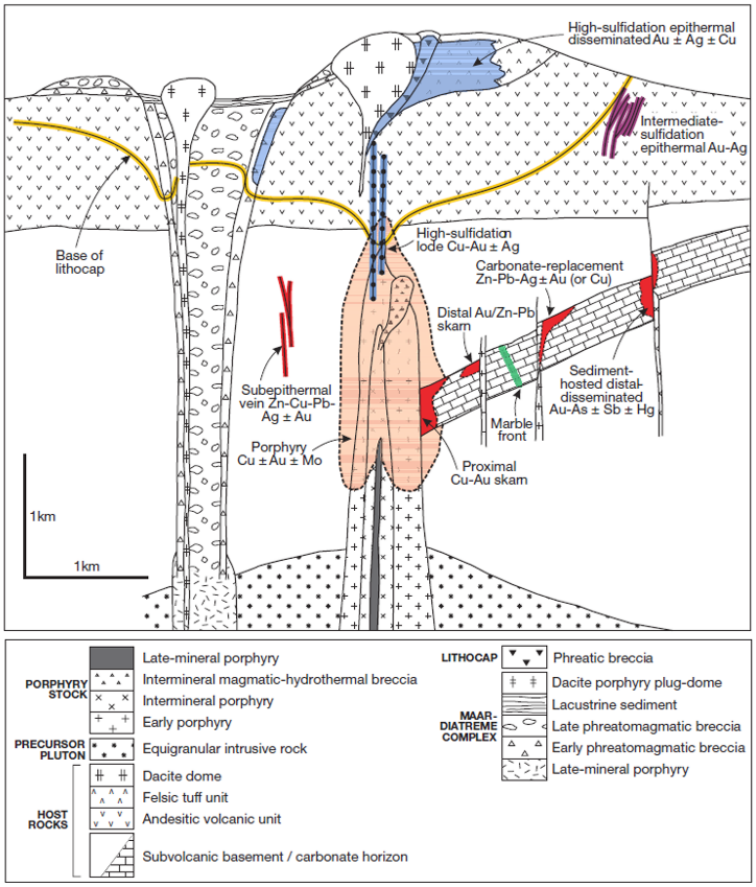


FIGURE 2.4: A generalised model of a porphyry system, showing the spatial relationship between porphyry, epithermal, skarn, carbonate replacement and sediment-hosted deposits (Sillitoe, 2010).

Section 2.2 of this review will focus primarily on those generalised characteristics, applicable idealised deposit models. The ore deposit models selected for review are those which fall within the scope of research comprising this thesis. While the formation and genetic characteristics are important to an overall understanding, this review will focus on the exploration model for each deposit style: quantifiable properties, how these have conventionally been considered as vectors to mineralisation and how they are predicted to manifest in the datasets used in our research in a manner amenable to a MLA approach.

2.2.1 Orogenic gold deposits

Orogenic gold deposits are named so due to a temporal association with the later stages of orogenic events (Groves et al., 2000). This deposit class is also known as lode gold, shear-hosted gold or mesothermal gold. As the name implies, this deposit class is sought after and mined for Au. The deposit class occurs throughout the Earth's Archaean cratons and continuously throughout Phanerozoic paleo-orogens (Goldfarb, Groves & Gardol, 2001). The following discussion will focus on the characteristics of the Archaean orogenic gold setting of Australia's Yilgarn craton pertaining to the orogenic Au component of our current research.

The majority of orogenic gold deposits in the Yilgarn are hosted in greenstone component of granite-greenstone terrains, named for the observed contrasting greenschist to lower amphibolite metamorphosed corridors of volcanic/intrusive and volcano-sedimentary rocks interspersed between granitic bodies (Goldfarb et al., 2001). Research has shown that such gold deposits favour greenschist to lower-amphibolite grade, Fe rich and structurally competent rock although a range of hosts are possible (e.g. Eilu & Groves 2001; Goldfarb et al., 2001; Groves et. al, 1998; Yeates & Vandehor, 1998; and sources cited therein). This metamorphic grade suggests that these deposits formed at depths between 10-20 km under a brittle to brittle-ductile regime, with pressure and temperature conditions important controls on the precipitation of gold.

Structure plays a key role in the formation of this deposit class with many examples occurring on second or higher order structures, connected to larger basement tapping structures. It is currently accepted that orogenic gold deposits are syn-tectonic, in that they occur during seismic events (or accumulate during a series of events), with structural controls on fluid flow, pressure and fluid rock interactions critical to gold precipitation (Cox & Ruming, 2004; Micklethwaite & Cox, 2006; Crawford, 2012). Deposits geometries are typically narrow and spatially discrete, ranging from several metres to tens of metres in width along the length/plane of host structures (Goldfarb et al. 2001).

Orogenic gold deposits are often surrounded by a hydrothermal alteration halo, zoned perpendicular to the plane of the host structure. This zonation is described by McCuaig & Kerrich (1998), as representing the progression of a metasomatic front of a fluid moving out from the centre of a deposit, eventually reaching equilibrium with the country rock. This alteration halo can extend for several hundred metres. The geochemical properties of this radially zoned alteration are shown in Figure 2.5.

It is acknowledged that direct detection of this deposit style is difficult (Yeates & Vandehor 1998; Anand & Butt, 2010) and as such proxy criteria are required. As structure places a key role in this style of deposit, the ability to map with sufficient accuracy and precision both regional structures and higher order structures is important. Doyle (1990), in a comprehensive review of geophysical exploration for gold, lists high resolution magnetics as one a key tool for mapping to this level of accuracy through the definition of magnetic units and subsequent mapping of offsets. In the 21st century, the collection of high resolution magnetic data is considered standard practice amongst gold explorers. Likewise, high resolution gravity can also aid in mapping these structures, albeit to a lesser extent, while also

having the ability detect low density regions concurrent with potentially favourable source intrusions (Doyle, 1992).

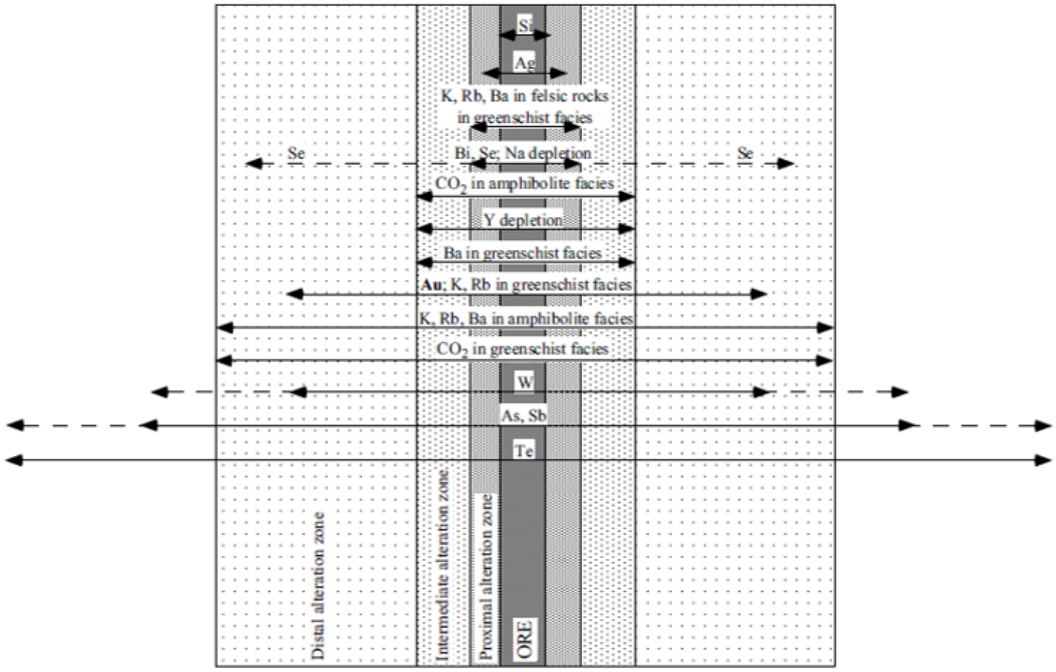


FIGURE 2.5: Generalised mineral and metal zonation associated with Archaean orogenic gold deposits (Eilu and Groves, 2001).

In addition, small-scale gravity surveys have been identified as important tools for identifying structures favourable to orogenic gold mineralisation due to the regolith cover in the Yilgarn. Anand & Butt (2010), published a comprehensive review and guide to exploration through the regolith in the Yilgarn. They discuss the need to understand the regolith as a means of exploring for gold at its base. They identify remote sensing tools including aerial photography, satellite and airborne multispectral imagery and radiometric surveys as a key to regolith mapping.

While it is possible for electrical and electromagnetic methods to detect orogenic gold deposits with substantial pyrite/pyrrhotite, the sulphide bearing zone is generally too narrow to produce an anomally when measured from surface or from the air (Smith 2014). Whitford, Meyers & Stolz (2005) deployed sub-audio magnetics to great effect at the St. Ives gold mine, in the Eastern Goldfields as a means of mapping depth changes in the regolith. These changes in regolith thickness highlighted preferential weathering along structures, however, these data are not widely available across the Yilgarn. The ability to detect pathfinder elements and minerals through geochemical sampling can provide important information on relative proximity to mineralisation. This can be directly sampled through whole rock geochemical techniques, or as suggested by Anand & Butt (2010), mapped via their dispersion patterns in the regolith via ground sampling or multispectral data.

As discussed earlier in this review, MLAs search for patterns in a higher dimensional space than traditional methods. This will allow for a combination of

the high resolution of magnetic and multispectral data for high spatial resolution and geochemical data for accurate discrimination of lithology.

2.2.2 Porphyry deposits

Porphyry deposits are defined as large, low grade, high tonnage mineral systems, produced by, and spatially related to porphyritic intrusions (Sillitoe, 2010). This deposit style can be further subdivided into classes or end-members based on their dominant economic metal content, namely Au, Cu or Mo (Sillitoe, 2010). For the purpose of this review, the generalised model, consistent to all end members will be discussed. The most influential model for porphyry systems, based on studies of the Kalamazoo deposit and still widely used today, was presented by Lowell and Guilbert (1970; Figure 2.5). Many studies have been conducted on porphyry deposits (e.g. Cooke et al. 1998; Gustafson & Hunt, 1975; Jones, 1992; Lowell & Guilbert, 1970; Shinohara and Hedenquist, 1997; Sillitoe, 2010). This research has refined porphyry models. Despite many local variations "per deposit", the mineralisation, alteration, tectonic setting, favourable host rocks and overarching geological, structural, chemical, and thermal mechanisms of ore formation are well defined.

The porphyry model described by Lowell & Guilbert (1970) comprises a porphyritic felsic to intermediate intrusion or stock, sourced from a larger, deeper batholith displaying radially zoned hydrothermal alteration assemblages. From the hotter, more central to cooler more distal parts of the system the alteration zonation is defined as follows: potassic, phyllic and propylitic, with variable amounts of advanced argillic overprinting. Mineralisation is generally associated with the potassic and to a lesser extent, phyllic alteration phase (Figure 2.6).

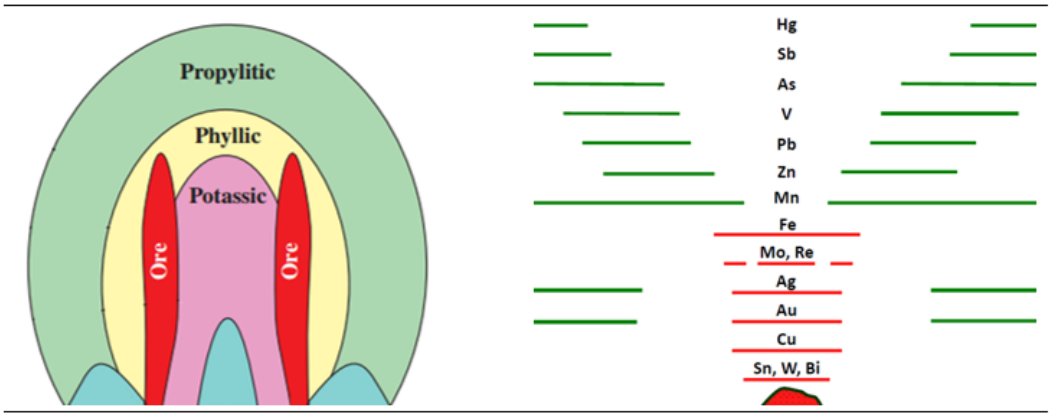


FIGURE 2.6: The simple porphyry model (modified from Berger et al., 2008 after Lowell and Guilbert, 1970). Right: Metal zonation associated with porphyry mineralisation from proximal (bottom and red) to distal (green; Cooke, 2014 after Jones, 1992).

Hydrothermal alteration associated with porphyry mineralisation comprises many measureable characteristics due to variable mineralogy and hence, elemental composition. Metal zonation from the inner to outer parts of the system provides a useful indication of location in a system and a vector towards mineralisation (Figure 2.5). This mineral and metal zonation can be mapped through

multi-element geochemical sampling. In addition, it has been shown that the enrichment or destruction of certain minerals can be expressed as geophysical anomalism. Previous research related to the magnetic properties of porphyry systems (e.g. Clark et al., 2004, 2014; Holden et al., 2011; Hoschke 2011) consistently identified a zone of low magnetic susceptibility around many porphyry systems, attributed to alteration halos favouring pyrite above magnetite. An oxidised zone, associated with potassic alteration and containing up to 3% magnetite, has been observed in Au-rich Porphyry systems (Hoschke, 2011). This may occur within the broader magnetic low, creating a "bullseye" effect, when viewed in plan. This contrast in magnetic susceptibility due to variable magnetite content manifests as zones of anomalously high and/or low total magnetic intensity (TMI) relative to the more homogenous background signal. Indeed, concentricity of magnetic anomalism, regardless of outright amplitude, can be a diagnostic exploration criterion for porphyry systems (Holden et al., 2011).

Elevated pyrite concentration (up to 10%), associated with the phyllic alteration zone, provide another target criterion, which is detectable by electrical geophysical methods (Hoschke, 2011). Induced polarisation can be a particularly effective technique in detecting disseminated pyrite in this zone (Hoschke, 2011; Cooke et al. 1998). Disseminated sulphides in the ore zone, such as chalcopyrite, and pyrite, in the phyllic zone, if sufficiently well connected, can behave as a massive body and subsequently be detectable by electromagnetic methods. The presence of conductive clay alteration minerals may also provide a detectable parameter by electrical and electromagnetic means. Areas of low density associated with granitiod batholiths to which porphyry deposits spatially adjacent to are detectable gravity data. Radiometric surveys can be used to map elevated K where the inner alteration zones present near enough to the surface and to map any enrichment in radiogenic elements K, U and Th that may be associated granitic intrusions (Hoschke, 2011; Cooke et al., 1998). Multispectral remote sensing may be useful in mapping lithology and mineral zonation where exposed at surface. Pour & Hashim (2011) demonstrated, through the use logical operators and shape-fitting based on partial mixing techniques applied to the shortwave infrared (SWIR) bands of ASTER (Abrams, 2000), the ability to map the alteration zones around porphyry deposits.

The presence of the above listed features in available data provides the foundation to accurate mapping and identification of porphyry systems. Correct identification of populations in the data indicative of alteration mineralogy and metals provide a means to identify position within a porphyry system and possibly a vector to mineralisation.

2.2.3 Epithermal deposits

Epithermal deposits were initially defined by Lindgren (1933). More recent classification of the deposit type, currently in use (Hedenquist et al., 2000), has subdivided epithermal deposits into three sub-types: low, intermediate and high sulphidation, based on sulphidation states of observable hypogene sulphide assemblages (see Table 2.1). Epithermal deposits are both spatially and genetically related to porphyry systems (Sillitoe 2010), as can be seen in Figure 2.4. The primary commodities extracted from epithermal deposits are Au and Ag.

Epithermal deposits occur in a range of terrains and rock types, although they to share common metal content (ratios vary by subtype) and alteration mineralogy. The distinction made by Hedenquist et al. (2000), based on observable sulphide characteristics, was used to facilitate classification in the field. The model for all subtypes is similar, a heat source (magma, often related to porphyry mineralisation) mobilises hot fluids towards the surface. Decompression boiling at temperatures between approximately 150-300^oC facilitates metal precipitation as the fluid moves through the system. Low and intermediate sulphidation deposits show evidence of meteoric-magmatic fluid mixing at depth. In contrast, high sulphidation deposits are believed to have formed as a result of magmatic fluids ascending and mixing with meteoric fluids higher in the system (White & Hedenquist, 1995). Epithermal deposits, due in part to their position at or near surface, exhibit a number of mapable properties, although this is dependant on the erosional level per deposit (Hedenquist, 2000).

TABLE 2.1: Generalised diagnostic features of epithermal Au-Ag deposits (from Sillitoe & Hedenquist, 2003)

	High sulfidation		Intermediate sulfidation	Low sulfidation	
	Oxidized magma	(Reduced magma)[1]		Subalkaline magma	Alkaline magma
Type example	El Indio, Chile (vein); Yanacocha, Peru (disseminated)	Potosí, Bolivia Fresnillo, Mexico (Ag-rich)	Baguio, Philippines (Au-rich);	Midas, Nevada	Emperor, Fiji
Genetically related volcanic rocks	Mainly andesite to rhyodacite	Rhyodacite	Principally andesite to rhyodacite but locally rhyolite	Basalt to rhyolite	Alkali basalt to trachyte
Key proximal alteration minerals	Quartz-alunite/APS; quartz-pyrophyllite/ dickite at depth	Quartz-alunite/APS; quartz-dickite at depth	Sericite; adularia generally uncommon	Illite/smectite-adularia	Roscoelite-illite-adularia
Silica gangue	Massive fine-grained silicification and vuggy residual quartz		Vein-filling crustiform and comb quartz	Vein-filling crustiform and colloform chalcedony and quartz; carbonate-replacement texture	Vein-filling crustiform and colloform chalcedony and quartz; quartz deficiency common in early stages
Carbonate gangue	Absent		Common, typically including manganiferous varieties	Present but typically minor and late	Abundant but not manganiferous
Other gangue	Barite common, typically late		Barite and manganiferous silicates present locally	Barite uncommon; fluorite present locally	Barite, celestite, and/or fluorite common locally
Sulfide abundance	10–90 vol %		5–>20 vol %	Typically <1–2 vol % (but up to 20 vol % where hosted by basalt)	2–10 vol %
Key sulfide species	Enargite, luzonite, famatinite, covellite	Acanthite, stibnite	Sphalerite, galena, tetrahedrite-tennantite, chalcopyrite	Minor to very minor arsenopyrite ± pyrrhotite; minor sphalerite, galena, tetrahedrite-tennantite, chalcopyrite	
Main metals	Au-Ag, Cu, As-Sb	Ag, Sb, Sn	Ag-Au, Zn, Pb, Cu	Au ± Ag	
Minor metals	Zn, Pb, Bi, W, Mo, Sn, Hg	Bi, W	Mo, As, Sb	Zn, Pb, Cu, Mo, As, Sb, Hg	
Te and Se species	Tellurides common; selenides present locally	None known but few data	Tellurides common locally; selenides uncommon	Selenides common; tellurides present locally	Tellurides abundant; selenides uncommon

Hoschke (2011), presented geophysical data for four epithermal deposits on the Pacific Rim as well as a review of prior research. Hoschke's definition of the geophysical exploration criterea was consistent with previous systematic reviews on the deposit style (e.g. Feebrey et al. 1998; Hedenquist et al., 2000; Sillitoe & Hedenquist, 2003). The afore mentioned study identified consistent geophysical anomalism associated with various parts of the ore system. Electrical methods in particular are useful for identifying components of epithermal deposits. The presence of silica associated with mineralisation is seen as a zone of high resistivity, while sulphides comprising mineralisation are often chargeable and observed in induced polarisation (IP) methods (Figure 2.7A; 2.7B). Likewise the presence of a lithocap can also manifest as a highly resistive but chargeable (if disseminated pyrite is present) anomally. A lithocap can also present as a topographic high,

associated with a radiometric low due to acid leaching. Hoschke noted that in many cases there was a broad zone of magnetite destruction associated with silica emplacement and an environment that favoures pyrite, noting however that a subdued response from the underlying units, in the case of high sulphidation systems, or linear anomalies confined to major structures within a broader magnetic low, in the case of low and intermediate sulphidation systems.

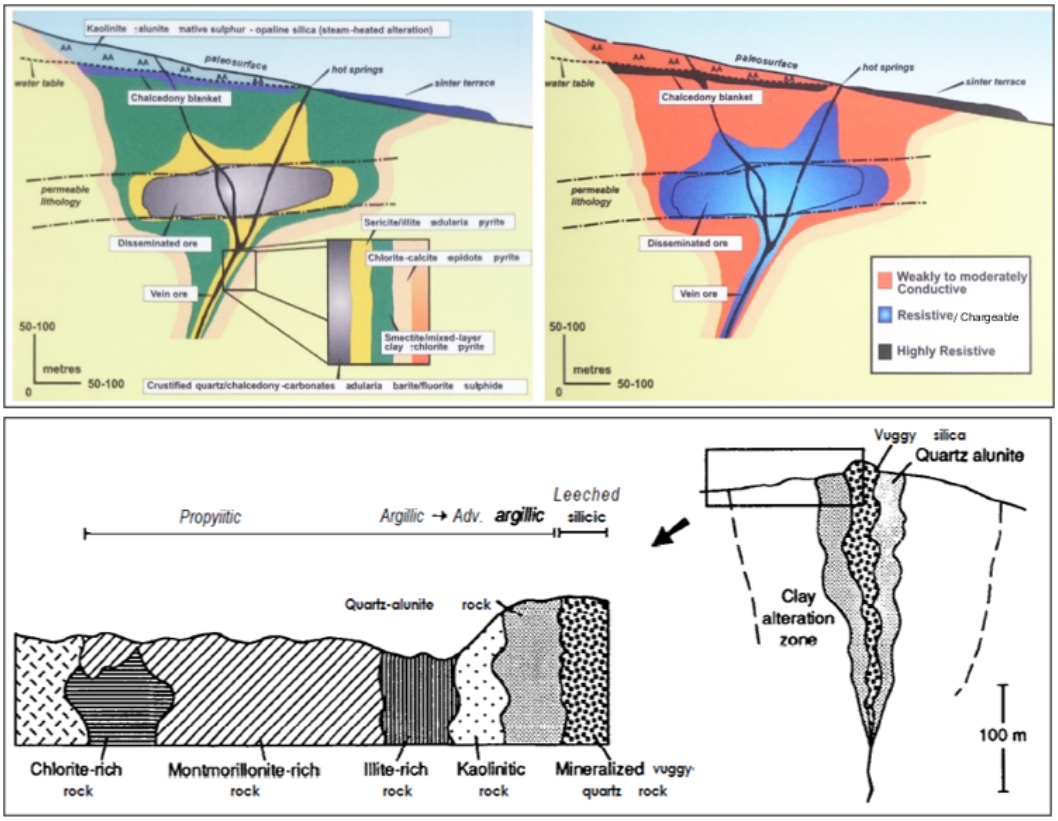


FIGURE 2.7: A (top): Alteration zonation associated with a low sulphidation epithermal system (left) and electrical properties, which also hold true for high and int. sulphidation systems (right) (after Hoschke, 2011). B (bottom): alteration zonation of a high sulphidation epithermal system from Hedenquist et al. (2000).

It was noted consistantly in the epithermal review papers listed above, that the gravity method is important in regional mapping but was innefective in pinpointing epithermal systems. In a study of the porphyry and epithermal systems in the Mankayan district of northern Luzon, Philippines, Chang et al. (2011) showed that SWIR techniques could aid in lithocap mapping. They observed a change in the alunite absorbtion peak towards higher wavelengths as sample locations approached the centre of a system. They attribute higher Na and reduced K in alunite to this phenomena. This suggests that SWIR can be an effective pathfinder in epithermal mineral systems and so by extension, multispectral imagery such as ASTER, which has 6 bands operating in the SWIR spectra may also prove to be beneficial (Abrams, 2000). In studies of whole rock geochemistry, using only non mineralised, alunite bearing samples, Chang et al. (2011) also observed an increase in Sr/Pb and La/Pb ratios and a decease in Pb, Hg, Ag and Ag/Au when moving outward from the centre of a system. While it is possible this may be idiosyncratic to the Mankayan/Far South East system, the results are promising for

mapping the system through a SWIR and systematic whole rock geochemical approach. This mode of data acquisition is favoured for producing regularised spatially contiguous input layers for inclusion in an RF/SOM style of assessment.

2.2.4 Sediment hosted copper systems of the Central African Copper Belt

Sediment hosted copper refers to a broad class of loosely stratabound (restricted to specific layers but not restricted to following bedding) copper deposits formed within sedimentary bedding (Cox, et al., 2007; Hayes et al., 2010). These systems are responsible for approximately 25% of the world's copper production, primarily sourced from the Central African Copper Belt and the Kupferschiefer of central Europe, while being an important source of cobalt and silver (Hitzman et al., 2010). Common characteristics of these deposits (Figure 2.8) are: 1) Oxidised source rocks such as red beds and basement rocks with sufficient mafic / ferromagnesian content from which to leach copper. 2) Oxidised brines, sourced from overlying evaporites, to scavenge copper from the source rocks and transport to 3) An appropriate trap site. The latter requires a sedimentary layer, generally argillite or arenite (Selley et al., 2005) with sufficient porosity and a reducing fluid to cause precipitation of copper. Reducing fluids can take the form of a hydrocarbon or be sources from organic rich shales and carbonaceous rocks which leads to these rocks being generally considered prospective for sediment hosted copper.

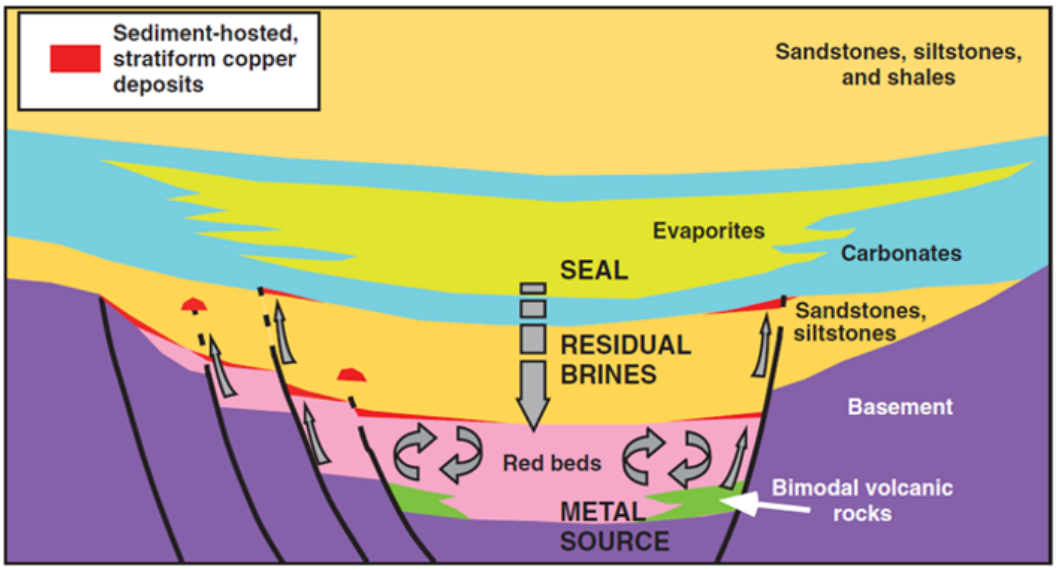


FIGURE 2.8: Schematic representation of the sediment hosted copper mineral system (from Hitzman et al., 2010)

These deposit types are associated with rifting (Brown, 1997), facilitating the deposition source and host stratigraphy and evaporites to supply the brines. Large scale basin growth structures provide a means of allowing copper enriched brines to move to and interact with reducing fluids at the site of deposition. Prolonged extension provides the necessary timeframe for these systems to become enriched (Hitzman et al., 2005). At some stage, a change in the hydrodynamic regime is required to force copper enriched fluids back up higher in the system where they will precipitate copper at the first reducing horizon encountered. This may be immediately above oxidised stratigraphy or via a structural path to a reduced unit

higher in the sequence. The action of salt can also facilitate fluid movement and interaction within the system (Hayes et al., 2010; Koziy et al., 2009). Hydrocarbons, in addition to providing a redox trap for copper bearing fluids, can aid in creating and maintaining porosity and permeability necessary for a deposit to form (Selley et al., 2005).

The characteristics of sediment hosted copper systems, as described above, allow mapping, characterisation and potentially, direct detection. This could be done either conventionally or using a machine learning approach, taking information from a number of geophysical, geological and geochemical properties (Table 2.2). Mapping of lithology can be approached using including magnetic, radiometric, gravity and electromagnetic data. These datasets can also be used to map structure and infer stratigraphic position where petrophysically distinct marker stratigraphy are known and present. Favourable host stratigraphy such as organic rich shales may be directly detectable as conductors and may have a magnetic response if sufficient pyrrhotite is present. It is possible to identify oxidised vs reduced regions in some cases due to the suppression of magnetic character in the oxidised regions where magnetite is oxidised to hematite. Geochemical data can be used to define lithology and stratigraphy via a lithogeochemical approach while a mineralised system may display anomalous Cu, Ag, Co, Pb, Zn, Mo, Re, V, Ge and U (Hayes, 2010). Accurate mapping of lithologies is key to determining stratigraphic position and allows normalisation of geochemistry to better map oxidised and reduced regions and the most probable location of mineralisation.

2.2.5 Overview of presented case studies

Individual studies that have investigated the use of RF are limited in number and restricted to specific settings. Much of the research cited in this review has focused on a particular training data scenario, either in the type of data used or variable spatial extent and availability of selected training data, e.g. Cracknell et al. (2014) which was specific to a VHMS setting. My research will broaden in the scope of use of RF to a variety of settings, terrains, training data paradigms representing a variety of exploration stages (Table 2.2). Previous research shows that there is opportunity to explore the methods used in untested geological settings and to undertake a more robust approach to comparing and rigorously testing training data and variable selection strategies directly in each deposit setting. An overview of the ore deposit styles on which my research will be focused indicates that there are characteristic features to each deposit style, which manifest in geophysical, geochemical and remote sensing data. Understanding these characteristics provides the foundation for the conversion of machine learning outputs into real geological knowledge.

TABLE 2.2: Summary of features characteristic of deposit styles reviewed in this chapter.

Deposit Style	Key Feature	Geophysical / Remote Sensing Response	Geochemistry / Pathfinders
Porphyry (Cu, Au, Mo)	Disseminated sulphide ore zone	Chargeable, magnetic (Au Systems), sometimes conductive	Biotite, K-feldspar (± magnetite, pyrite, chalcopyrite)
	Potassic alteration halo	Magnetic high (Au Rich Systems)	Biotite, K-feldspar (± magnetite)
	Phyllic alteration halo	Magnetic low, chargeable, sometimes conductive	Quartz, sericite, pyrite
	Metal zonation	Elevated Fe associated with Pyrite (chargeable) and/or magnetite (magnetic high)	Zonation from centre: Sn, W, Bi, Cu, Au, Mo, Fe, Mn, Zn, Pb, V, As, Sb, Hg
	Source batholith	Gravity low, Elevated K, U, Th, magnetic low (sometimes)	-
	Multiphase intrusions	Low chargeability relative to mineralisation (sometimes)	Geochemistry may allow for distinction of intrusions vs hydrothermal alteration
	Advanced argillic alteration halo	Lithocap may be observable in aerial photography and multi-spectral imagery	Kaolinite, quartz, alunite, pyrophylite, dickite
Epithermal High Sulphidation (HS), Low Sulphidation (LS) (Au-Ag)	Sulphide mineralisation	Chargeable response at range of depths. Gradient array useful for surface coverage	**HS:** Au-Ag, Cu, As, Sb (minor Zn, Pb, Bi, W, Mo, Sn, Hg), **IS:** Ag-Au, Zn, Pb, Cu (minor Mo, As, Sb), **LS:** Au, Ag (Minor Zn, Pb, Cu, Mo, As, Sb, Hg
	Silica associated with mineralisation	Highly resistive (broad: HS, Narrow: LS - gradient array good for surface coverage), can be topographically prominent. Subdued magnetic response	-
	Clay alteration around mineralisation	magnetite destructive (broad: HS, narrow: LS)	SWIR or whole rock, mineral identification (e.g. illite, alunite, pyrophyllite)
	Adularia-Illite alteration (IS, LS)	Radiometrics, elevated K	-
	Deep intrusion	May manifest in gravity response if there is a density contrast with host rocks (often not)	-
	Alteration zonation reflecting temperature gradient	Can be mapped using SWIR (e.g. ASTER mineral maps) where exposed	Zoned: biotite, mica (boiling); illite, smectite, smectite, kaolinite (whole rock geochem, portable SWIR)
	Lithocap	Topographically prominent, radiometric low associated with intense leaching, resistive	SWIR Alunite 1480nm peak, Na/ (Na + K)

Table 2.2 - Continued

Deposit Style	Key Feature	Geophysics/ Remote Sensing	Geochemistry/Pathfinders
Archaean Orogenic Gold (Au)	Basement tapping structure	Gravity gradient associated domain change, large offset of magnetic units, change in magnetic texture across structure	Change in geochemistry associated with juxtaposed rock units
	Higher order local structures	Offset magnetic units, high conductivity zones associated with deeper, preferential weathering along structures	-
	Favourable host	Magnetic character of host rock	Primary geochemistry of unaltered host rock (varies per host rock)
	Radial metasomatic alteration zonation	Can control redox gradient - magnetite formation or destruction, detectable via magnetic survey (sometimes)	Varies per host rock and metamorphic grade - refer to Figure 2.7
	Discrete zone of disseminated sulphides in ore zone	Discrete - can be locally chargeable and conductive (rarely detectable at larger distances)	-
	Age of mineralisation/ favourable stratigraphy	Possible known remnant magnetisation in any rock units present	Primary geochemistry of host rock to determine stratigraphic unit, not just rock type
	Regolith properties	Discrimination of lithology and possible alteration through radiometric (K, Th, U) and multispectral imagery (Landsat, ASTER, SWIR)	Regolith soil sampling/SWIR, geochemical signatures corresponding to underlying fresh rock lithology.
Sedimentary Copper (Cu) (Central African Copper Belt Deposits)	Lithology and favourable stratigraphy	Mapping geology via magnetics, gravity, EM, radiometrics and remote sensing. Host shales often conductive (EM)	Identification of lithology and stratigraphic position via lithogeochemistry / immobile trace elements
	Fluid pathways	Mapping of faults in magnetic and electromagnetic data	Lithogeochmistry. juxtaposition of units at fault contacts, unconformities etc.
	Empirical association with granite domes (African Copper Belt)	Mapable in magnetic and gravity data (gravity low)	Directly observable in mapping and lithogeochemistry where at or near surface
	Reduced vs oxidised lithologies	More complex magnetic character in reduced areas, subdued response in oxidised basement rocks (magnetite destruction)	Reduced stratigraphic units enriched in V, Mo, U
	Ore, alteration, and pathfinders	Ore sulphides conductive, may be seen in magnetics if sufficient pyrrhotite	Cu, Ag, Co, Pb, Zn, Mo, Re, V, U, Co, Ge

References

(n.d.).

Abrams, M. (2000). The advanced spaceborne thermal emission and reflection radiometer (aster.

Aitchinson, J. (1982). The statistical analysis of compositional data, **44**(2).

Anand, R. and Butt, C. (2010). A guide for mineral exploration through the regolith in the Yilgarn craton, Western Australia, *Australian Journal of Earth Sciences* **57**(8): 1015–1114.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding.

Bedini, E. (2009). Mapping lithology of the Sarfartoq carbonatite complex, southern West Greenland, using hymap imaging spectrometer data, *Remote Sensing of Environment* **113**(6): 1208–1219.

Berger, B., Ayuso, R., Wynn, J. and Seal, R. (2008). Preliminary model of porphyry copper deposits, *U.s. Geological Survey open-file report 2008–1321.*

Bierlein, F., , Fraser, S., Brown, W. and Lees, T. (2008). Advanced methodologies for the analysis of databases of mineral deposits and major faults, *Australian Journal of Earth Sciences* **55**: 79–99.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

Breiman, L. (2001). Random Forests, *Machine Learning* **45**: 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth International Group, Pacific Grove, California.

Briggs, I. (1974). Machine contouring using minimum curvature surfaces, *Geophysics* **39**: 39–48.

Brown, A. (1997). World-class sediment-hosted stratiform copper deposits: Characteristics, genetic concepts and metallotects, *Australian Journal of Earth Sciences* **44, 3**: 317–328.

Chang, Z., Hedenquist, J., White, N., Cooke, D., Roach, M., Deyell, C., Garcia, J., Gemmell, J., Mcknight, S. and Cuison, A. (2011). Exploration tools for linked porphyry and epithermal deposits: Example from the Mankayan intrusion-centered cu-au district, *Luzon, Philippines; Economic Geology* **106**: 1365–1398.

Chiles, J. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty, 2nd edition.*, Wiley Series in Probability and Statistics.

Clark, D. (2014). Magnetic effects of hydrothermal alteration in porphyry copper and iron-oxide copper–gold systems: A review, *Tectonophysics* **625**(0): 46–65.

Cooke, D., Heithersay, P., Wolfe, R. and A, C. (1998). Australian and western Pacific porphyry cu-au deposits, *ASGO Journal of Australian Geology and Geophysics* **17**(4): 97–104.

Cox, D., Lindsay, D., Singer, D., Moring, B. and Diggles, M. (2007). Sediment-hosted copper deposits of the world: Deposit models and database, *U.s. geological survey open-file report 03-107*.

Cox, L. A. (1982). Artificial uncertainty in risk analysis, *Risk Analysis* **2**: 121–135.

Cox, M. and Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization, *Proceedings of the 8th conference on Visualization* **97**: 235–244.

Cox, S. and Rumming, K. (2004). The St Ives mesothermal gold system, Western Australia — a case of golden aftershocks?, *Journal of Structural Geology* **26**: 1109–1125.

Cracknell, M. (2014). *Machine Learning for geological mapping: Algorithms and applications.*, University of Tasmania.

Cracknell, M. and Reading, A. (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): 113 – 126.

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

Cracknell, M., Reading, A. and De Caritat, P. (2015). Multiple influences on regolith characteristics from continental-scale geophysical and mineralogical remote sensing data using Self-Organizing Maps, *Remote Sensing of Environment* **165**: 86–99.

Cracknell, M., Reading, A. and McNeill, A. (2013). Supervised and unsupervised classification of near-mine soil geochemistry and geophysics data, *ASEG - PESA 23rd International Geophysical Conference and Exhibition*, Melbourne.

Cracknell, M., Reading, A. and McNeill, A. (2014). Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests^{TM} and Self-Organising Maps, *Australian Journal of Earth Sciences* **61**: 287–304.

Crawford, M. (2011). *Dynamic Coupling Between Deformation Processes, Fluid-Rock Interaction, and Gold Deposition in the Argo Gold Deposit*, St Ives, Western Australia; Australian National University.

Cutler, D., Edwards, T., Beard, K., Cutler, A. and Hess, K. (2007). Forests for classification in ecology.

Davidson, G. and Large, R. (1998). Proterozoic copper-gold deposits, *ASGO Journal of Australian Geology Geophysics* **17**(4): 105–113.

Doyle, H. (1990). Geophysical exploration for gold - a review, *Geophysics* **55**(2): 134–146.

Dutton, D. and Conroy, G. (1996). A review of machine learning, *The Knowledge Engineering Review* **12**(4): 341–367.

Eilu, P. and Groves, D. (2001). Primary alteration and geochemical dispersion haloes of Archaean orogenic gold deposits in the Yilgarn Craton: the pre-weathering scenario, *Geochemistry: exploration, environment, analysis* **1**(3): 183–200.

Farr, T. G., R. P. A. C. E. C. R. D. R. H. S. K. M. P. M. R. E. and Roth, L. (2007). The shuttle radar topography mission.

Feebrey, C., Hishida, H., Yoshioka, K. and Nakayama, K. (1998). Geophysical expression of low sulphidation epithermal au-ag deposits and exploration implications -examples from the Hokusatsu region of SW Kyushu, Japan.

Fraser, S. and Dickson, B. (2008). A new method for data integration and integrated data interpretation: Self-Organising Maps.

Goldfarb, R., Groves, D. and Gardoll, B. (2001). Orogenic gold and geologic time: a global synthesis, *Ore Geology Reviews* **18**: 1–75.

Goodchild, M., Sun, G. and Shiren, Y. (1992). Development and test of an error model for categorical data, **6**: 87–104.

Goorvaerts, P. (1997). *Geostatistics for Natural Resource Evaluation*, Oxford University Press.

Groves, D. (2010). Iron oxide copper-gold (IOCG) deposits through earth history: Implications for origin, lithospheric setting, and distinction from other epigenetic iron oxide deposits, *Economic Geology* **105**(3): 641–654.

Groves, D., Goldfarb, R., Gebre-Mariam, M., Hagemann, S. and Robert, F. (1998). Orogenic gold deposits: A proposed classification in the context of their crustal distribution and relationship to other gold deposit types, *Ore Geology Reviews* **v. 13**: 7–27.

Gustafson, L. and Hunt, J. (1975). The porphyry copper deposit at El Salvador, Chile, *Economic Geology* **70, 5**: 857–912.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

Hayes, T., Cox, D., Piatak, N. and Seal, R. (2015). Sediment-hosted stratabound copper deposit model, *U. S. Geological Survey scientific investigations report 2010–5070–m*.

Hedenquist, J., Arribas, R. and Gonzales-Urien, E. (2000). Exploration for epithermal gold deposits, *Reviews in Economic Geology; Society of Economic Geologists* p. 245–277.

Hitzman, M., Kirkham, R., Broughton, D., Thorson, J. and Selley, D. (2005). The sediment-hosted stratiform copper ore system, *Economic Geology* **100**: 649–642.

Hitzman, M., Oreskes, N. and Einaudi, M. (1992). Geological characteristics and tectonic setting of proterozoic iron oxide (Cu, U, Au, REE) deposits, *Precambrian Research* **58**(1–4): 241–287.

Hitzman, M., Selley, D. and Bull, S. (2010). Formation of sedimentary rock-hosted stratiform copper deposits through Earth history, *Economic Geology* **105, 3**: 627–639.

Hitzman, M. and Valenta, R. (2005). Uranium in iron oxide-copper-gold (IOCG) systems, *Economic Geology* **100**(8): 1657–1661.

Holden, E., Fu, S., Kovesi, P., Dentith, M. and Bourne, B. (2011). Automatic identification of responses from porphyry intrusive systems within magnetic data using image analysis, *Journal of Applied Geophysics* **74**(4): 255–262.

Hood, S., Cracknell, M. and Gazley, M. (2018). Linking protolith rocks to altered equivalents by combining unsupervised and supervised machine learning, *Journal of Geochemical Exploration* **186**: 270–280.

Hoschke, T. (2011). *Geophysical signatures of copper-gold porphyry and epithermal gold deposits, and implications for exploration; ARC Centre of Execellence in Ore Deposits*, University of Tasmania.

Jones, B. (1992). Application of metal zoning to gold exploration in porphyry copper systems, *Journal of Geochemical Exploration* **43**(2): 127–155.

Klose, C. (2006). Self-Organizing Maps for geoscientific data analysis: geological interpretation of multidimensional geophysical data, *Computational Geosciences* **10**(3): 265–277.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics* **v. 43**: 56–69.

Kohonen, T. (2001). *Self-organizing maps*, Springer series in information sciences, 30, Springer-Verlag, Berlin.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques; informatica (Ljubljana), *ASEG - PESA 23rd International Geophysical Conference and Exhibition*.

Koziey, L., Bull, S., Large, R. and Selley, D. (2009). Salt as a fluid driver, and basement as a metal source, for stratiform sediment-hosted copper deposits, *Geology* **37, 12**: 1107–1110.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C. and Engelhardt, A. (2012). Classification and Regression Training, R package version 5.15-023,.
URL: *http://CRAN.R-project.org/package=caret*

Leung, Y., Goodchild, F. and Lin, C. (1993). Visualization of fuzzy scenes and probability fields, *Computing Science and Statistics* p. 416–422.

Lindgren, W. (1933). *Mineral deposits*, McGraw-Hill, New York.

Lloyd, S. (1957). Least squares quantization in pcm: Technical note; Bell Laboratories, *Published in IEEE Transactions on Information Theory* **28**(2): 129.

Lowell, J., D., G. and J., M. (1970). Lateral and vertical alteration - mineral zoning in porphyry ore deposits, *Economic Geology* **65**: 373–408.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, University of California Press, Berkeley, p. 281–297.

Mann, J. (1993). *Uncertainty in geology, Computers in Geology - 25 Years of progress*, Oxford University Press, Oxford.

Mars, J. and Rowan, L. (2006). Regional mapping of phyllic- and argillic-altered rocks in the Zagros magmatic arc, Iran, using Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) data and logical operator algorithms, *Geosphere* **2**(3): 161–186.

McCuaig, T. and Kerrich, R. (1998). P-T-t-deformation-fluid characteristics of lode gold deposits: Evidence from alteration systematics, *Ore Geology Reviews* **12**(6): 381–453.

Micklethwaite, S. and Cox, S. (2006). Progressive fault triggering and fluid flow in aftershock domains: Examples from mineralised Archaean fault systems, *Earth and Planetary Science Letters* **250**: 318–330.

Mitchell, T. (1997). *Machine Learning*, McGraw-Hill, New York.

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. and Duchesnay, (2011). Scikit-learn: Machine learning in python, *Journal of Machine Learning Research* **12**: 2825–2830.

Pour, A. and Hashim, M. (2012). The application of aster remote sensing data to porphyry copper and epithermal gold deposits, *Ore Geology Reviews* **44**(0): 1–9.

Rodriguez-Galiano, V., Chica-Olmo, M. and Chica-Riva, s. M. (2014). Predictive modelling of gold potential with the integration of multisource information based on Random Forest: a case study on the Rodalquilar area, *Southern Spain; International Journal of Geographical Information Science* **28**(7): 1336–1354.

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

Selley, D., Broughton, D., Scott, R., Hitzman, M., Bull, S., Large, R., McGoldrick, P., Croaker, M. and Pollington, Y. (2005). A new look at the geology of the Zambian Copperbelt, *Economic Geology* pp. Hundredth Anniversary Volume, 965–1000.

Shannon, C. (1948). A mathematical theory of communication, *Bell Systems Technical Journal* **27**: 379–423.

Sillitoe, R. (2010). Porphyry copper systems, *Economic Geology* **v. 105**: 3–41.

Sillitoe, R. and Hedenquist, J. (2003). Linkages between volcanotectonic settings, ore-fluid compositions and epithermal precious metal deposits; volcanic, geothermal and ore-forming fluids: Rulers and witnesses of processes within the earth. 1 ed.: Johnson Printing.

Skirrow, R., Bastrakov, E., Barovich, K., Fraser, G., Creaser, R., Fanning, C., Raymond, O. and Davidson, G. (2007). Timing of iron oxide Cu-Au-(U) hydrothermal activity and nd isotope constraints on metal sources in the Gawler craton, South Australia, *Economic Geology* **102**: 1441–1470.

Smith, R. (2014). Electromagnetic induction methods in mining geophysics from 2008 to 2012, *Surveys in Geophysics* **35**(1): 123.

Tayebi, M. and Tangestani, M. (2015). Sub pixel mapping of alteration minerals using SOM neural network model and hyperion data, *Earth Science Informatics* **8**(2): 279–291.

Turing, A. (1950). Computing machinery and intelligence, *Mind* **v. 59**: 433–460 , 433–460.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer-Verlag New York, Inc.

Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley Sons Inc.

Waske, B., Benediktsson, J., Árnason, K., and Sveinsson, J. (2009). Mapping of hyperspectral aviris data using machine-learning algorithms, *Canadian Journal of Remote Sensing* **35**: 106–16.

Wellmann, J. and Regenauer-Lieb, K. (2012). Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models, *Tectonophysics* **529**(0): 207–216.

White, N. and Hedenquist, J. (1995). Epithermal gold deposits: Styles, characteristics and exploration, *SEG Newsletter* **23**: 1–13.

Whitford, M., Meyers, J. and Stolz, N. (2005). The SAM EQMMR response of the regolith at East Victory, St Ives Gold Mine, Western Australia, *Exploration Geophysics* **36**(- 2): – 139.

Williams, P., Barton, M., Johnson, D., Fontbote, L., De Haller, A., Mark, G., N., O. and Marschick, R. (2005). Iron oxide copper-gold deposits: Geology, space-time distribution, and possible modes of origin, *Economic Geology* (Anniversary Volume): 371–405.

Yeats, C. and Vanderhor, F. (1998). Archaean lode gold deposits, *ASGO Journal of Australian Geology and Geophysics* **17**(4): 253–258.

Yu, L., Porwal, A., Holden, E., J. and Dentith, M. (2012). Towards automatic lithological classification from remote sensing data using support vector machines, *Computers and Geosciences* **45**(0): 229–239.

Chapter 3

Lithological mapping using Random Forests applied to geophysical and remote sensing data: a demonstration study from the Eastern Goldfields of Australia

Stephen Kuhn[12], Matthew J. Cracknell[12] and Anya M. Reading[23]

[1] Centre for Ore Deposit and Earth Sciences (CODES) University of Tasmania.
[2] ARC Industrial Transformation Research Hub for Transforming the Mining Value Chain (TMVC), University of Tasmania.
[3] School of Natural Sciences (Maths and Physics), University of Tasmania.

3.1 Abstract

The Eastern Goldfields of Western Australia is one of the world's premier gold producing regions, however, large areas of prospective bedrock are under cover and lack detailed lithological mapping. Away from the near-mine environment exploration for new gold prospects requires mapping geology using the limited data available with robust estimates of uncertainty. In this study, we use the machine learning algorithm, Random Forests, to classify the lithology of an underexplored area adjacent to the historically significant Junction gold mine, using geophysical and remote sensing data, with no geochemical sampling available at this reconnaissance stage. Using a sparse training sample, 1.6% of the total ground area, we produce a refined lithological map. The classification is stable, despite including parts of the study area with later intrusions and variable cover depth, and preserves the stratigraphic units defined in the training data. We assess the uncertainty associated with this new Random Forests classification using information entropy, identifying those areas of the refined map which are most likely to be incorrectly classified. We find that information entropy, correlates well with inaccuracy, providing a mechanism for explorers to direct future expenditure towards areas most likely to be incorrectly mapped or geologically complex. We conclude that the method can be an effective additional tool available to geoscientists in a greenfields, orogenic gold setting, when confronted with limited data. We demonstrate that the method could be used either to improve substantially an existing map, or produce a new map, taking sparse observations as a starting point. It can be implemented in similar situations (with limited outcrop

information and no geochemical data) as an objective, data driven alternative to conventional interpretation with the additional value of quantifying uncertainty.

3.2 Introduction

With increasing cost and difficulty of new discovery in areas with substantial amounts of cover, there is a need for improved approaches to mineral exploration. In the Eastern Goldfields of Australia, very detailed geological, geochemical and geophysical datasets exist near mines. There is, however, a sharp transition into adjacent greenfields areas where such data are not available and the geology is significantly less-well constrained. Geophysical and remote sensing data are widely available at a reasonable resolution either in the form of government or multi-client datasets; or as a first pass acquisition performed by explorers when new ground is acquired. Machine learning presents an attractive way forward, facilitating the use of these data to improve a preliminary lithology map or to produce a starting map from limited observations: in each case improving an explorer's ability to identify targets. Previous studies however (e.g. Cracknell et al., 2014, Harris and Grunsky, 2015; Waske, 2009) have primarily used a richer and more diverse set of data inputs such as geochemistry or additional spectral information or, made use of a different algorithm, such as, for example, support vector machines (SVM; Yu et al., 2012) or Artificial Neural Networks (Barnett and Williams, 2009). In this study we assess the ability of the machine learning algorithm (MLA) Random Forests (RF) to produce a geological classification using only those geophysical and remote sensing data that would be available to an explorer in a greenfields, early stage exploration environment.

3.2.1 Geological setting

The Heron South project area is located approximately 15 km east of the Junction gold mine, in the St Ives Goldfield of the Yilgarn Craton, Western Australia (Figure 3.1, Figure 3.2). The St Ives camp is estimated to contain in excess of 300 t of gold, with orogenic and (to a lesser extent) intrusion related gold deposits hosted throughout the entire local stratigraphy, making it one of Australia's largest gold producing districts (Crawford, 2011). The Archean (2.7-2.6 Ma) bedrock stratigraphy comprises a series of mafic-ultramafic volcanic and intrusive units, volcanoclastic sediments and felsic intrusions, cross-cut by Proterozoic aged basaltic dykes. The region has undergone pervasive, regional greenschist to lower amphibolite metamorphism. The St Ives Goldfield is bound to the west and east by the Merougil and Boulder-Lefroy Fault Zones respectively.

The region was subject to several distinct phases of deformation between 2675 and 2620 Ma. Until recently, the deformational framework for the region had largely focused on compressional events (e.g. Swager, 1997, Ngyuen, 1997, Connors et al., 2002). The more recent study by Blewett et al. (2010) includes events related to extension, important in understanding the formation of the younger volcano-sedimentary units of the region (Squire et al. 2010). The revised framework proposed by Blewett et al. (2010) is as follows: D_1 is characterised by ENE-WSW extension. D_2 represents a phase of ENE-WSW contraction that caused regional NNW-trending folds and re-activation of faults produced during D_1 as thrusts. This was followed by D_3, a period of extension on the same orientation. D_{4a} was a

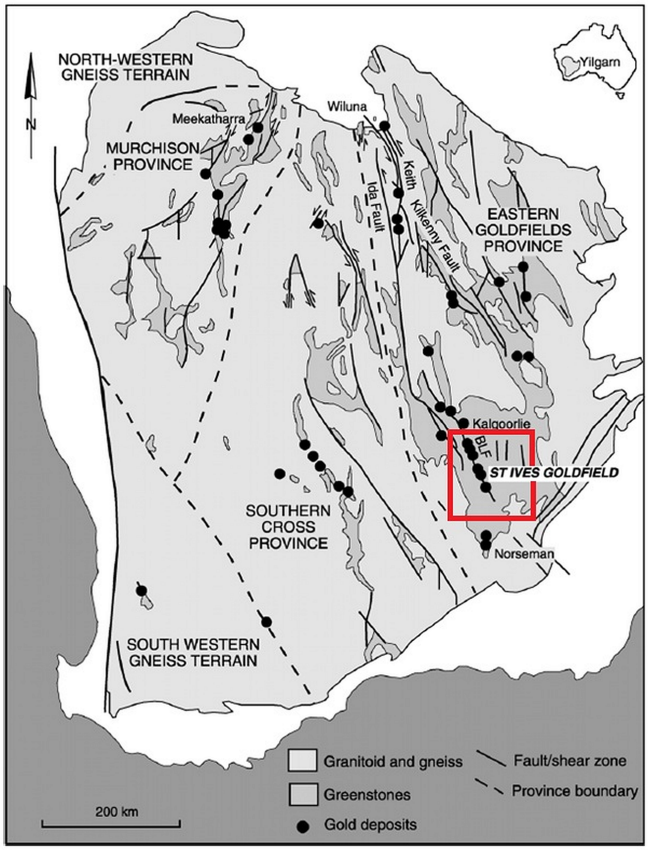


FIGURE 3.1: Schematic representation of the Yilgarn Craton including the location of major gold deposits. The approximate location of the Heron South project is shown in red (modified after Cox and Ruming, 2003).

period of contraction that tightened existing NNW folds and was followed by D_{4b}, a period of sinistral transpression. During this deformation event existing structures, such as the Boulder Lefroy Fault Zone, which passes through Heron South, were reactivated as sinistral strike slip faults. Localised deflections, step-overs and local higher order structures produced during D_{4b} are associated with the main mineralising event in the region (Blewett et al., 2010, Cox and Ruming, 2004, Miller et al., 2010). D_5 was a period of dextral transtension producing NNE trending strike slip high angle faults. These structures may also be associated with a gold mineralising event at various sites in the region (Blewett et al., 2010, Ruming, 2006, Miller et al., 2010, Connors et al., 2002). D_6 is not documented in the St Ives Goldfield. D_7 was a period of contraction associated with the emplacement of dominantly ENE trending Proterozoic dykes which occur in abundance in the study area. The Heron South project area is proximal to the Boulder-Lefroy Fault Zone which passes through the southwest of the project on a NNW orientation.

The geology of the study area is split by the Boulder Lefroy Fault Zone into a western, and an eastern region. The western area forms part of the main St Ives sequences and contains thick successions of Paringa Basalt and Black Flag Group volcano-sedimentary sequences. The eastern area contains N-S striking, steeply dipping packages of mafic-ultramafic and sedimentary units impinged between

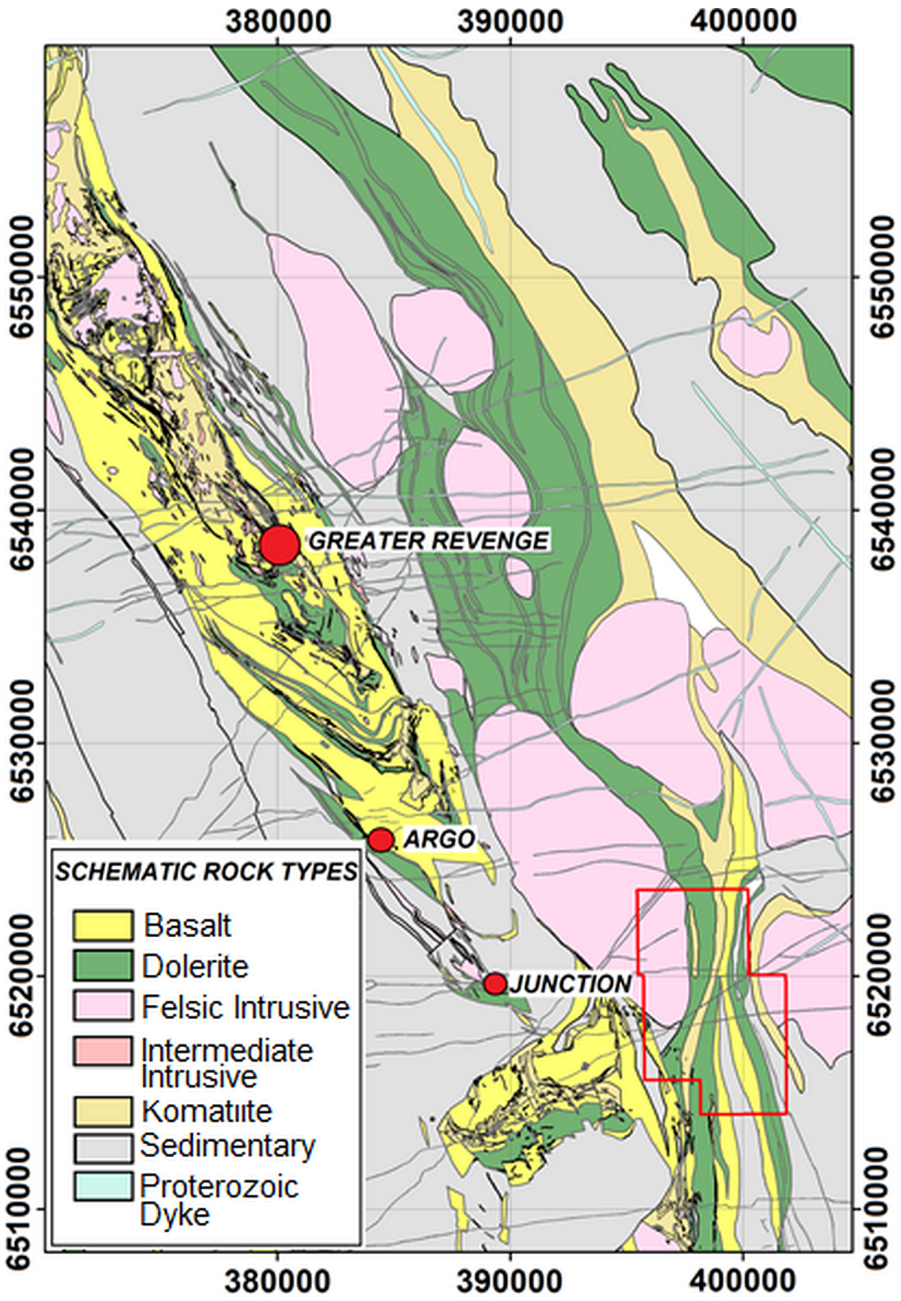


FIGURE 3.2: Schematic geology of the St. Ives gold camp with including the location and extent of the project (red outline box) relative to several major existing and historical gold mines (indicated by red circles with mine name adjacent). The project outline (red box) defines the extent of the project in all subsequent figures. Map coordinates are projected using WGS84, UTM grid 51S (m).

larger granitoid bodies (Figure 3.2). It is anticipated that these units are correlates of the main stratigraphic sequence mapped at St Ives; however, this has not yet been confirmed. For the purpose of this study, these units have been defined by the interpreted geological map of the St Ives Goldfield (Figure 3.3), as stratigraphically distinct. Stratigraphic labels can be assigned to these units as geochemical and geochronological information becomes available allowing these units to be amalgamated or subdivided as required at a future date.

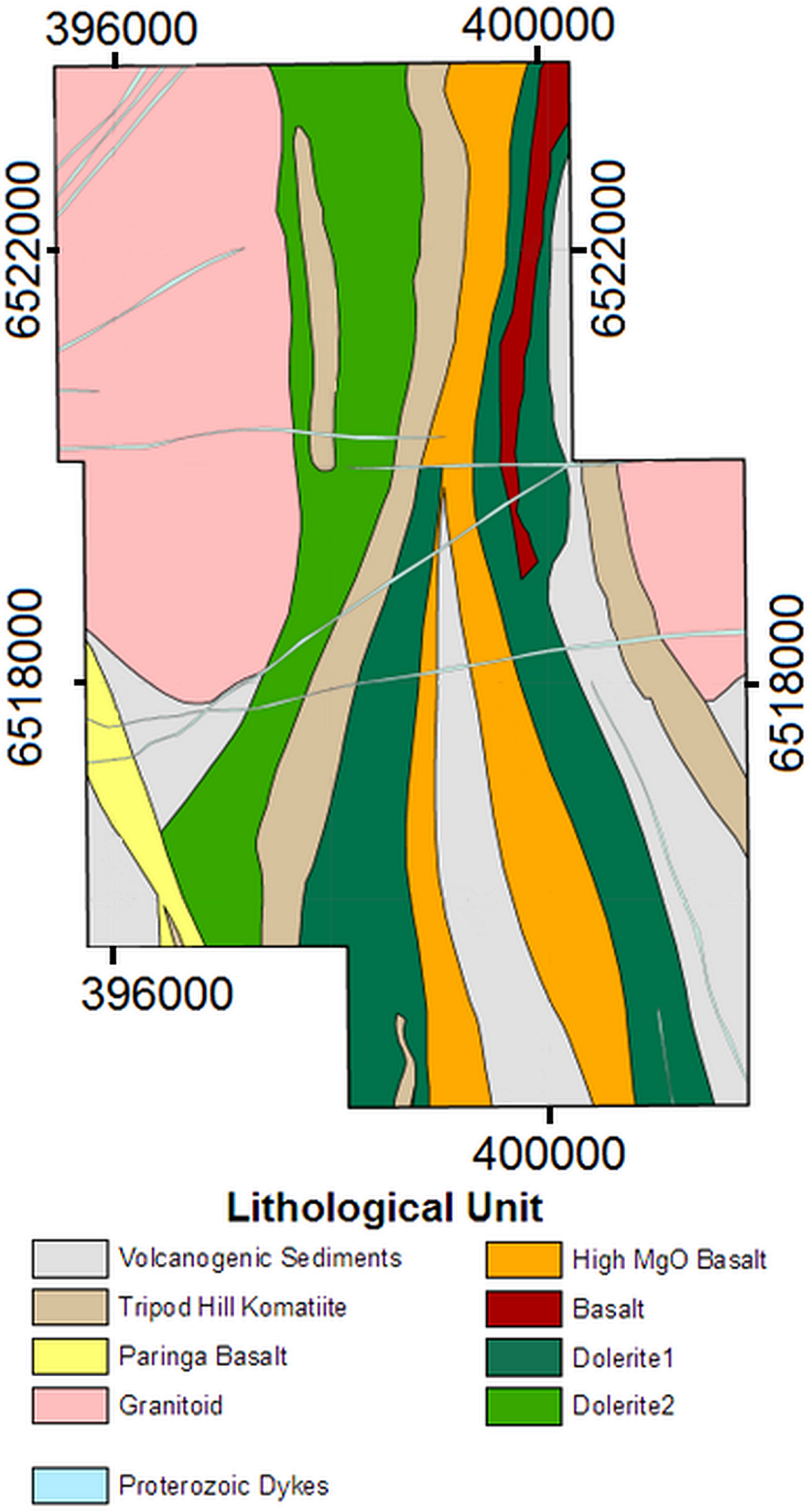


FIGURE 3.3: Heron South geology map. In subsequent figures, the lithological units will be abbreviated as follows: Volcanogenic Sediments (VS), Tripod Hill Komatiite (THK), Paringa Basalt (PB), Granitoid (G), High MgO Basalt (HMgOB), Basalt (B), Dolerite 1 (D1), Dolerite 2 (D2). The map extent in this figure defines the extent of all subsequent map figures in this article.

3.2.2 Random Forests

RF (Breiman, 2001, p. 5-32) is a supervised ensemble classification algorithm and an extension of the decision trees method. This classifier constructs a 'forest' comprising many decision trees (Figure 3.4), allowing for superior performance and lower sensitivity to over fitting compared to single classifiers (Hastie et al., 2009, p. 587-604). Randomness is introduced at two stages during implementation of the algorithm. Firstly, a process of bootstrap aggregation, known as bagging (Breiman, 1996 p. 123-124) is used to modulate the training data (Ta) available to each decision tree. Bagging obtains for each tree, via random sampling with replacement, a subset of Ta equal in size to Ta. This duplicates some samples and will not select others. An average of 63.2 % of instances are included in each training subset, while the remaining or "out-of-bag" samples (37.8%) are used for validation. The second form of randomisation involves the selection of variables available to the classifier to split each node. At each node, a random subset of input variables selected from all available input variables. The number of variables in this subset is predefined and consistent across the forest. At every node, the randomly selected variables are then ranked by ability to produce a split threshold that maximises the homogeneity of child nodes (Figure 3.4) relative to the parent node. The decrease in Gini index (Equation 1), as implemented by Breiman et al. (1984) provides this measure. The Gini index is an expression of information purity given by:

$$\text{Gini(t)} = \sum_{c=1}^{j} g_c(1 - g_c)$$

where g_c is an expression of the relative frequency of each class c, of a set comprising j classes, at a given node t. g_c is given by:

$$g_c = \frac{n_c}{n}$$

where n_c is the number of samples comprising class c at a given node and n is the total number of samples comprising that node. Using this measure, the variable which produces the greatest improvement in homogeneity in child nodes relative to the parent node is used to split the node at the threshold which produced the best split. This is repeated at every node until sufficient depth is reached to produce nodes with complete homogeneity (or approached to within a defined tolerance). The class assigned by RF to each sample is determined by a majority vote compiled from the output of all classification trees (Breiman, 2001, p. 6).

Many studies have noted a point of diminishing returns, necessitating a forest be grown to a certain extent where a stable error minima is approached, beyond which, additional trees are redundant (e.g. Cracknell et al., 2014; Harris and Grunsky, 2015; Rodriguez-Galiano, Chica-Olmo and Chica-Rivas, 2014; Waske, 2009). RF has been shown to achieve equal or better accuracy to other classification algorithms with the advantage that parameter selection is relatively straightforward. (e.g. Cracknell and Reading 2013; Hastie et al., 2009). process of RF training can be performed on any PC with specifications readily commercially available at the time of this study and does not require specialised equipment. In this study, combined training and cross validation of a RF for any given set of parameters required between 15 and 40 seconds on a Dell Precision T7610 with an Intel Xeon e2630 processor and 32Gb RAM. This is ideal for uptake by geoscientists as requirements for specialised computing skills and equipment are minimal.

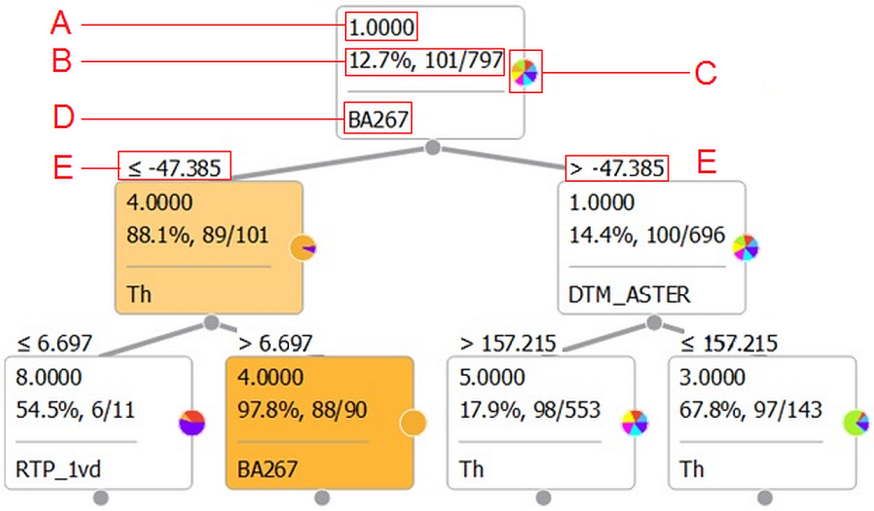


FIGURE 3.4: An example showing 3 levels of a classification tree, showing at each node: A) the most numerous class, B) the proportion of samples of the most numerous class, relative to all samples in the node (shown as percentage and count of total), C) pie graph distribution of all classes present, D) variable used to split parent node into child nodes and E) the threshold at which that split was executed.

RF has been increasingly applied to the problem of lithological classification. Waske et al. (2009), compared RF and another popular MLA, Support Vector Machines (Vapnik, 1995, 1998), in the context of mapping lithology using hyperspectral imagery. They concluded that both RF and SVM achieved significantly more accurate results than standard classifiers. While in that instance, SVM marginally outperformed RF, it was noted by the authors that RF remained an attractive option due to high accuracy and ease of use. Cracknell and Reading (2014) compared RF with four other MLAs: SVM, Naïve Bayes, k-Nearest Neighbours and Artificial Neural Networks; as applied in to lithological mapping. In their study, RF marginally outperformed other MLAs. While there were only small differences in accuracy, Cracknell and Reading (2014) demonstrated that RF was able to produce accurate results with simpler input parameters and at less computational cost than other algorithms evaluated. Another study by Cracknell and Reading (2013) assessed RF and SVM for lithology mapping; and identification of lithological contacts and zones of structural complexity. They discovered that RF, in addition to an excellent overall performance, produced more usable outputs. Unlike for SVMs, high uncertainty was spatially associated with incorrect classification; and proximal to geological boundaries and zones of high structural complexity. Cracknell and Reading (2014) noted that with increasingly spatially dispersed training data, the comparative performance of RF improved further, widening the gap over other MLAs.

Cracknell and Reading (2014) demonstrated that RF was able to identify and redefine incorrectly mapped features in western Tasmania using 2 percent of the surface area as training samples. Harris and Grunsky (2015) utilised a similar approach, applying RF to geological mapping in northern Canada. They tested two Ta selection scenarios: one based on lake sediment geochemical sample locations

and another based on field mapping observations. Both approaches produced meaningful results with the authors concluding that RF is of value as a first-pass mapping tool or as a means of focussing effort into areas where there is a mismatch between predicted geology and legacy maps.

3.2.3 Information entropy

There has been an increasing effort in the field of mineral exploration to quantify the uncertainty associated with mapping and prediction. One such method, information entropy (H) (Shannon, 1948) is defined as:

$$H = -k \sum_{i=1}^{n} p_i log p_i$$

p_i is the class membership probability at location i, n is the number of candidate classes, k is an arbitrary positive constant. Both k and the base of the logarithm can be selected by the user to define scale. H has been used to great effect in a "per-voxel" setting to demonstrate how uncertainty is distributed spatially (Wellmann and Regenauer-Lieb, 2012). In the process of producing a final classification, RF calculates class membership probabilities. These are defined as the proportion of trees in a RF which voted for a given candidate class (Hastie et al., 2009). RF class membership probabilities can be used in Equation 3 to calculate H for each classified instance. The properties of H for a two class, binary, system are such that a value of 0 corresponds to a 100% probability of one class occurring and a value of 1 corresponds to an equal probability of both represented classes being present. H in its general form preserves monotonicity such that an increase in the number of candidate classes results in higher H. For the purpose of this study, a normalised version of H has also been used, to account for number of candidate classes by dividing H by the logarithm of the number of classes present, such that H assigned to each pixel represents, on a scale of 0–1, the range of minimum to maximum possible H for that pixel. As such, all pixels are comparable with regard to how close they each internally approach their minimum or maximum possible H. For example, a pixel with two possible and equally probable classes; and a pixel with four possible and equally probable classes; shall both be described as H being equal to 1.

3.3 Methods

3.3.1 Data

In this study 16 geophysical and remote sensing datasets were used (Figure 3.5), and interpolated at a grid cell size appropriate (20–25%) to their respective acquisition line spacing (Table 3.1). Landsat thematic mapper and Shuttle Radar Topography Mission (SRTM) products (United States Geological Survey, 2003) were procured in raster format and their original point separation specifications were preserved (National Aeronautics and Space Administration, 2006 and United States Geological Survey, 2003, respectively). Each dataset was re-sampled to a 30 metre grid in order to populate a matrix where each line takes the form of: x, y, p_1, p_2, ..., p_n , where x and y are spatial coordinates and p are the various measured properties at each pixel. At the extent of the study area, this comprised approximately 56,000 samples. The compiled data were split into subsets comprising training (T_a) and test (T_b) data through a process of stratified spatially

random sampling. 100 samples were taken from each of the eight lithological classes comprising the study area. These 800 samples comprising T_a, represent approximately 1.6% of the total dataset (Figure 3.6). The remaining 98.4% of data, T_b, were not shown to the classifier during the training process.

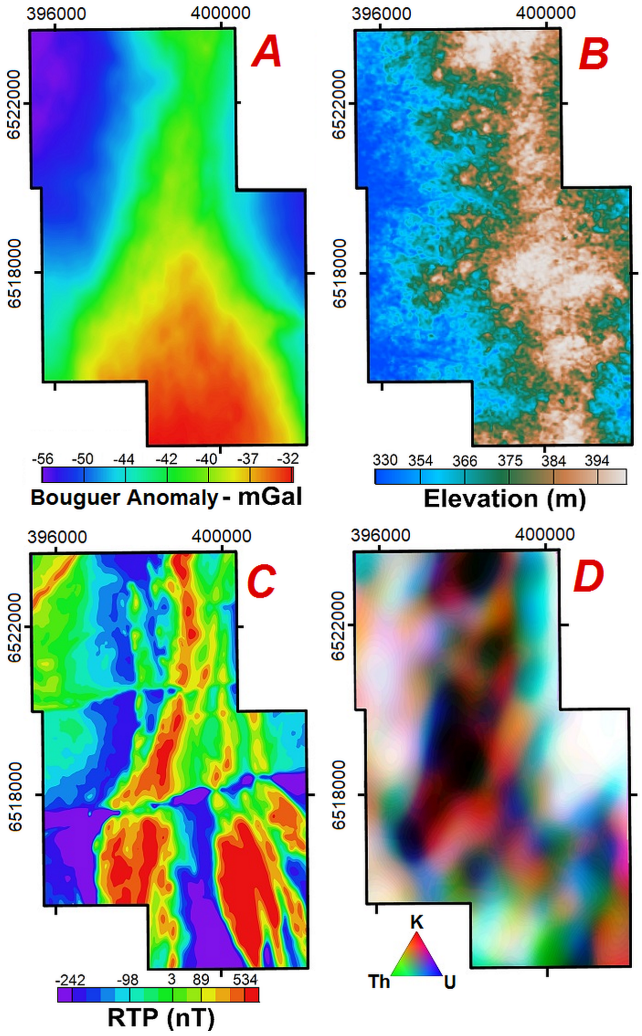


FIGURE 3.5: Examples of input data; A. Bouguer anomaly, B. Elevation, C. Reduced to pole total magnetic intensity and D. Ternary radiometric image.

3.3.2 Variable ranking and selection

RF facilitates several means of ranking the importance of input variables. In this instance each variable was permuted and the effect on out-of-bag classification accuracy was measured. Those variables which, when permuted, produced the greatest change to classification accuracy were ranked highest (Table 3.2.). Due to the relatively small number of datasets used in this study, none of the starting input variables were sufficiently well correlated (as defined by a threshold at a Pearson's correlation coefficient = 0.85) with one another to warrant removal, due to duplication of information, prior to ranking. To optimise both speed and interpretability of results, redundant variables were screened at this stage. Using Ta, variables were successively added to the classification according to their ranked

TABLE 3.1: Geophysical and remote sensing datasets used in study, including abbreviations and spatial resolution.

Dataset	Abbreviation	Spacing
Gravity (Bouguer Anomaly)	BA267	200 m x 200 m
1st Vertical Derivative of Gravity	BA267_1vd	200 m x 200 m
Airborne Magnetics – Reduced to Pole	RTP	50 m EW flight lines
1st Vertical Derivative of Airborne Magnetics	RTP_1vd	50 m EW flight lines
Airborne Mass Spectrometry: Potassium	K	50 m EW flight lines
Airborne Mass Spectrometry: Thorium	Th	50 m EW flight lines
Airborne Mass Spectrometry: Uranium	U	50 m EW flight lines
Elevation (Digital Terrain Model)	SRTM	90 m pixel
Landsat Thematic Mapper: Channels 1-8	LSb1-8	30 m pixel

importance established in the prior step. Accuracy was assessed using a forest comprising 500 classification trees, via 10-fold cross validation (Table 3.2). Cross validation accuracy improved with the input of additional variables, albeit at a diminishing rate, until a peak cross validation accuracy of 79% was achieved via the inclusion of variables ranked one to eight (Table 3.2). Beyond this point, no increase in cross validation accuracy was observed through inclusion of additional variables, as such, the Landsat data, ranked 9th to 15th, were omitted. This is logical given the sensitivity of reflectance methods to the immediate surface in an area heavily influenced by transported cover. Easting and Northing were omitted at this stage to avoid over fitting to classification based on position.

TABLE 3.2: Variable importance rankings as determined by RF and cross validation accuracy (CV Acc). Cross validation accuracy indicates the accuracy achieved when the corresponding variable is added in addition to higher ranked variables. Abbreviations are as per Table 3.1. Bold text indicates the first occurrence of peak cross validation accuracy corresponding to variables selected for classification.

Variable	Score RF	Rank	CV Acc (%)
BA267	20.61	1	35.1
Th	16.54	2	62.1
DTM	14.37	3	69
RTP	12.14	4	74.5
U	7.83	5	76.6
Ba267_1VD	7.18	6	78.4
K	2.69	7	79.5
RTP_1VD	1.52	**8**	**79.8**
LSb3	0.97	9	79.7
LSb7	0.69	10	79.1
LSb4	0.61	11	79
LSb5	0.58	12	79.8
LSb2	0.37	13	79.6
LSb8	0.26	14	79.8
LSb1	0.24	15	79.5
LSb6	0.19	16	79.4

3.3.3 Classification and uncertainty

800 samples comprising 100 from each of the lithological units defined above (Figure 3.6) were used to train a RF classifier. Each sample was attributed with the 8 non-redundant variables identified during variable ranking. We used a RF comprising 500 trees with no limits on individual tree depth or subsequent pruning. The RF produced under these parameters required 12 seconds to train. Subsequently, the remaining data comprising T_b, which do not have an associated class, were shown to the trained classifier and a class prediction for each was made. Class membership probabilities, describing the proportion of trees voting for each class, were retained for the calculation and assessment of H.

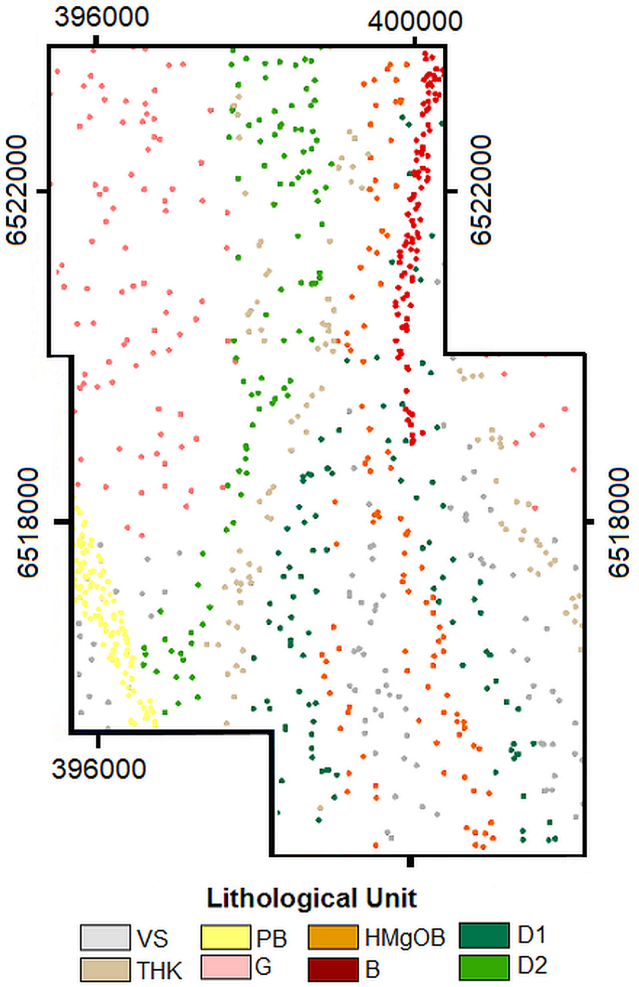


FIGURE 3.6: Ta location coded by lithology. Note that sample point diameter has been enlarged by a factor of 5 for legibility. Legend abbreviations are as described in Figure 3.3.

3.4 Results

RF produced a new version of the geological map (Figure 3.7A), correctly predicting mapped geology in 76.8% of Tb instances. The remainder of samples can be categorised as either incorrect predictions or as showing new information not previously mapped; or incorrectly mapped in the starting product. When plotted,

class probabilities produced by RF (e.g. Figure 3.7B–7D), show the spatial distribution of lithology dependent class membership probabilities. Areas where a class has a very high probability of occupying an area with little likelihood of another class being present such as, for example, the central zone of D2, (Figure 3.7C) are apparent. There are however, regions where multiple classes compete such that the class that ultimately is predicted displays a marginally higher probability than its competition (e.g. Figure 3.7B and 7D).

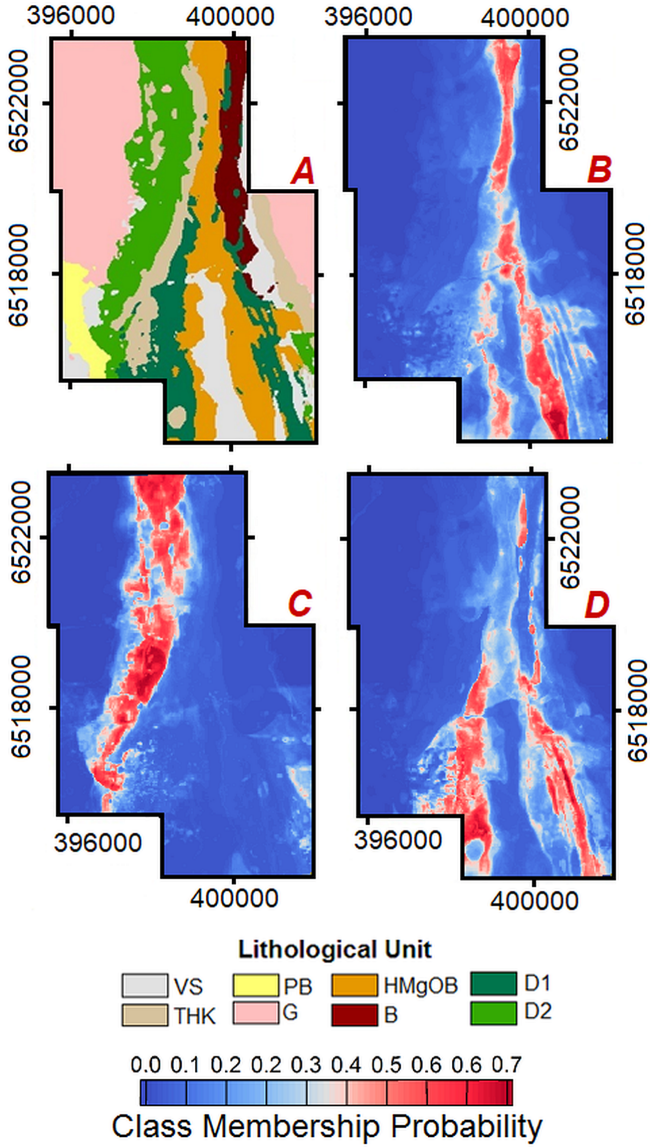


FIGURE 3.7: A) RF predicted Heron South geology. B) Probability of THK class, C) Probability of D2 class and D) Probability of D1 class. All class membership probabilities are presented on the same linear scale, shown at bottom of image. Lithology abbreviations are described in Figure 3.3.

The confusion matrix in Table 3.3 indicates, on a per class basis, the distribution of correct and incorrect classification percentages with respect to all other classes. Several classes, namely the basaltic and granitic units, have been predicted with a high degree of accuracy. One of the doleritic units (D2) is commonly classified by RF as basalt or high MgO basalt. This suggests that either the classification was

incorrect in this instance or alternatively, areas mapped as dolerite are in fact basalt. There is spatial control on classification accuracy with misclassification more likely when units with similar petrophysical properties occur adjacent to one another. The overlapping petrophysical signals of these classes, particularly in the case of potential field data due to smooth transitions as opposed to sharp boundaries, may be contributing to a reduced ability to make accurate predictions. This is particularly notable where these classes occupy the same areas of the map suggesting both similarity of properties and spatial proximity are factors.

TABLE 3.3: Confusion matrix comparing mapped class with RF predictions. Values are shown as a percentage of the number of samples of a class present in the interpretation map. Red, yellow and blue text indicates a recall greater than 50%, 70% and 80% respectively.

		Predicted							
		VS	THK	PB	G	HMgOB	B	D1	D2
Observed	VS	59	3.9	5	3.6	13	2.6	3.6	9.4
	THK	1.8	71.5	0.3	0.9	7.5	0.2	8	9.8
	PB	0.9	0	98.7	0	0	0	0	0.4
	G	1.2	0.4	0.6	89.8	0	0	0	8.1
	HMgOB	7.3	0	0	0	85.4	2.2	5.1	0
	B	0	0.2	0	3.6	0	95.7	0.5	0
	D1	5.8	3.4	0	0.2	14.9	16.5	59.1	0
	D2	0.8	10.5	2	2.2	0.1	0	2.7	81.8

The spatial distribution of H (calculated using Equation 3) shows very few examples where a candidate class has a 0 probability of occurrence in a given pixel. By definition this means that it must be included as a term in the calculation of H, mitigating the ability to display the monotonic increase in H that additional possible classes impose. As such a threshold probability of 2% was selected, below which a class can be considered, for this purpose, to be not present in that pixel. The calculation of H with this parameter imposed was used to produce a map of the spatial distribution of H (Figure 3.8A). Areas in the central north and southwest of the project display the highest H, indicating that these areas are characterised by a high level of uncertainty across multiple classes that display a relatively high probability of being predicted. Conversely, areas in the east and west of the project extent that are classified as granite coincide with low H, indicating that, RF classifications can be treated with a high degree of confidence such that no other classes have a high probability of being present. H when normalised for number of possible classes, represents the relative minimum to maximum possible H on a per-pixel basis (Fig 3.8B). There is a direct relationship between normalised H and the observed discrepancies between the interpretation map and that produced by RF. This correlation can qualitatively observed in a visual comparison of Figures 8B and 8D; and was confirmed quantitatively by Kuhn et al. (2016), who demonstrated statistically distinct populations of H corresponding to correctly and incorrectly classified sample groups. Both H and normalised H can potentially form the basis of the assessment of the quality of RF predictions in the absence of a starting map with which to compare.

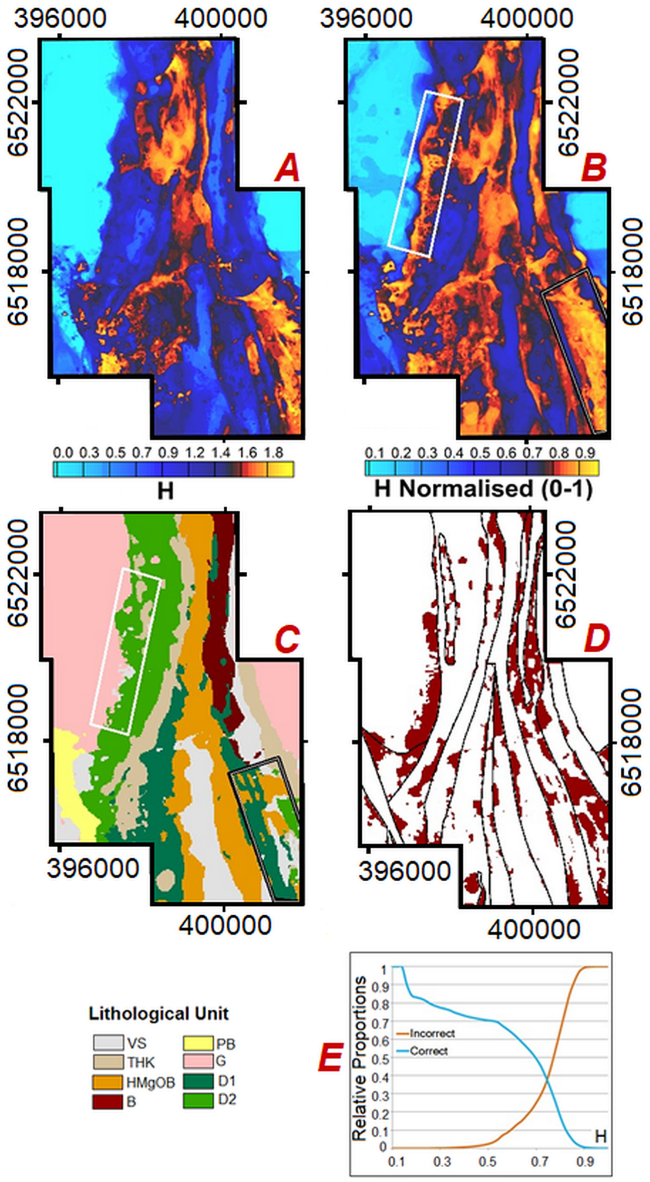


FIGURE 3.8: A) H (information entropy). B) H normalised per pixel to 0-1. C) Lithology predictions made by RF. D) Accuracy relative to starting map (white = correct, red = incorrect). E) The relative proportion of correctly (blue) and incorrectly (orange) classified samples (blue) at a given threshold of H. White box (8B and 8C) indicates a westward extension of D1 predicted by RF and associated high H increasing towards, and peaking at the geological boundary. White-black outlined box (8B and 8C) indicates a zone of potential geological complexity associated with high H. Figure 3.8 B C D and E modified from Kuhn et al., 2016.

3.5 Discussion

In the absence of the information that indicates orogenic gold mineralisation directly, the ability to map and interpret geology accurately is a key feature in target identification and the establishment of priority areas for exploration. We have demonstrated in this study that RF was able to classify lithology with an accuracy of approximately 76% relative to an existing interpreted geological map using only 2% of available data as training samples. These results are comparable to those achieved by Cracknell, Reading and McNeill (2014) who used a similar approach, achieving 78% accuracy, and compare favourably to similar implementations using SVM such as Yu et al. (2012) who achieved a consistency with the geology map of between 50.5% and 62.2% with various modal convolution filters applied. It is important however, to note that different data and geological conditions were encountered in each case. Nevertheless, the results of this study compare well with similar applications in different settings.

Looking beyond bulk similarities, there is a wide range in performance with regards to predictive power of the RF as applied to individual classes. As shown in Table 3.3, both the VS and D1 classes produced accuracies with respect to the starting geological map, in the order of 59% while the PB class exceeded 98%. It is likely that this excellent result is due to the spatially discrete and small area defined by the PB class, resulting in a very well constrained class signature. The poor performance of the VS class is likely due to a highly variable class signature, the result of both a wide range of sample locations and potentially, misidentification in the original map. D1 was commonly confused with B (16.5% of instances) and HMgOB (14.9%) which is logical, given the compositional similarity of these mafic units. D2 however, while quite accurately captured at 81.8% was confused most commonly with the THK class at a rate of 10.5% indicating the possibility of unmapped ultramafic material interspersed in the region mapped as D1, or conversely, doleritic intrusions in the THK. Alternatively, this could indicate erroneous mapping of these units in the original geological interpretation map. The G class was most often confused with the D2 class. This is explained by erroneous mapping in the starting map being re-partitioned into the D2 class which RF extends further to the west, supported by the expression of H in that region (Figure 3.8C).

In making use of spatially stratified random sampling, we have used a near ideal spatial sample of the project area. A more spatially or numerically imbalanced Ta would produce a less robust result. This study is, in part, an exercise in cross validation against an existing interpretation map and as such a training subset was taken at random. In a deployment of the method over incompletely mapped regions, the distribution of training data would be determined entirely by the number and positions of available geological observations, such as those obtained from outcrop or drilling. It is important however that, even when using geophysical data in the absence of geochemistry or mineralogical information, only a very small percentage of a project area need be observed provided these sample distribution criteria are met.

An important component of these results was the observation that RF was able to preserve class labels defined from stratigraphic relationships and distinguish between equivalent lithologies. In this case, the stratigraphic sequence is not well

characterised and geochemical data were not available to resolve this distinction. Geological interpretations indicate that multiple dolerites and basalts are present in this region. The contrast between greenstone, felsic to intermediate intrusive bodies and sedimentary packages is well expressed in the gravity and magnetic datasets facilitating mapping using these variables via machine learning. It is, however, difficult to distinguish between units of similar composition using these datasets alone. Nevertheless, RF is able to capture this distinction, to the extent that it was present in the training data, and produce a map retaining stratigraphy and not simply amalgamating by rock type. Results produced by RF do not indicate a large scale revision to the mapping or understanding of structure in the area. Updates to lithological boundaries could form the basis of an adjustment to the position of faults sub-parallel to stratigraphy and those which offset stratigraphy. Knowledge of the position within the stratigraphic column is important in an exploration context given that several models for the stratigraphic position of favourable host units, relative to the timing of gold deposition, have been identified. Again, this is contingent on the congruency of the sampled region. We suggest that when using geophysical data, the accuracy of RF lithological predictions cannot be assumed to apply to adjacent terrains. Potential field data in particular are influenced by effects such as cover depth or the response of deeper sources can produce a shift in absolute signal amplitude, not related to geology as mapped at surface. As such, the rules defined by RF are only reliably applicable to the domain and from which they were derived. Radiometric data are indicative of surficial features and may be mirrored in adjacent or distant domains however it is also likely that these data may be influenced by weathering and vegetation which differs from the study area. In any event, it is not anticipated that radiometric data alone would be sufficient to propagate mapping to greater distances beyond the sampled region. Our approach is designed as a pragmatic workflow, however, further insights might be gained by more geostatistical or computer-science oriented practitioners (e.g. Grunsky and Kjaarsgard, 2016).

It is important to note that regardless of the physical response, elevation, depth to source or height of sensor of a method, RF will preferentially use whichever variables allow the algorithm to most accurately solve the given problem, in this case, lithology. The datasets which are ranked highest, and the associated frequency response are entirely determined by which allow RF to discriminate between the lithologies.

The topographic (SRTM) dataset ranked highly amongst the available input data. Given the contiguity and dominant strike of the geology relative to topography, it is possible that topography is, in fact, serving as a proxy for lithological position in the landscape. It is also probable that rock composition is one of the controlling factors in preferential weathering and hence topography, although this relationship is not always obvious in the region.

The Bouguer anomaly and reduced to pole total magnetic intensity (RTP) datasets were both ranked as more important to the classification than their first vertical derivatives. The most plausible interpretation of this result being that the potential field data are more closely related to rock composition at the scale of this study. The respective derivatives may define detailed features of the units which could reflect structural or compositional variability. This information is of immense value in accurately mapping and interpreting the regional and within-unit structural

complexity of the area but does not necessitate a change to the lithological class at any given location. Should the mapping area be expanded, the effects of regional trends would become more significant with derivatives, as a form of high pass filter, being required to mitigate the influence of these trends and thus would likely be ranked of higher importance. It is possible that the introduction of additional, textural data, derived from those datasets could have improved results. It is worth noting however that of key importance is the ease of use of the method by geoscientists and as such we consider this a good demonstration of the method using readily available datasets, accessible to most projects without additional prerequisite knowledge of GIS operations.

H provided an indication of those areas where an operator can be confident of accurate mapping and those areas where they are more likely to be incorrect. Consistent with prior research (Cracknell and Reading 2013), high uncertainty was generally observed in proximity to lithological boundaries and areas of geological complexity. Kuhn et al. (2016) have demonstrated statistically, through examination of the distribution of normalised H in correctly and incorrectly classified (Figure 3.8F) samples that H provides a good, albeit imperfect proxy for inaccuracy. As such H is a valuable tool when mapping in unknown areas and where validation against a known result is not possible. Performing any exploration activity requiring fiscal expenditure through a decision unknowingly underpinned by a Type II statistical error in classification has a greater consequence than performing additional study on an area that in fact was mapped correctly. H highlights areas that require additional data collection, such that geoscientists can further validate these areas to within the scope of reasonable due diligence prior to additional expenditure. Conversely areas producing low H do not require the same level of attention and as such, effort need not be expended here and can be diverted to those areas of higher uncertainty. We believe that H is therefore a valuable mechanism for quantifying uncertainty given that in addition to a normalised product, the purest form of H preserves monotonicity and provides a measure of the absolute uncertainty present throughout the classification.

The presence of highly magnetic Proterozoic dykes often confounds the ability to interpret Archaean stratigraphy. A manual interpreter may opt to attempt to see past these features in a somewhat subjective manner. It does, however, prohibit the use of absolute levels in classification of individual datasets, such as aeromagnetic imagery, when analysing only that property, dykes are indistinguishable from other mafic units on a pixel by pixel basis. In this instance, our randomly selected training data included several samples of various rock units in the locations where they were intruded by Proterozoic dykes. As this interaction was represented in the training data, RF was able to map consistently the underlying geological class and was largely immune to the presence of these features. Looking at H, we can see that uncertainty does consistently increase by up to approximately 20 percent (Figure 3.8B) in in areas where dykes intrude other lithologies, however, the correct decision has still been obtained.

It is assumed that classifications produced by RF are deemed incorrect in the event that they do not conform to the geological map. An interpreted geological map, however, is a constantly evolving product. Both the accuracy and level of detail of an interpreted geological map improves as data of higher resolution and accuracy, and better interpretation techniques, become available. In a greenfields setting, where

a geological map is based on limited outcrop and interpretation of potential field datasets, it is entirely plausible that it contains errors and/or oversimplifications.

When RF produces a result which differs from the geological map training data is sourced from, H provides a means to assert whether the RF output or the reference information are likely to be incorrect. In this instance (Figure 3.8E), we can see that the western boundary of the greenstone package is moved to the west relative to its position in the interpreted map. Low H at the original boundary suggests RF predicted with high certainty that this was in fact an area of greenstone. In addition, H increases towards the predicted contact suggesting greater uncertainty as the transition between rock types was approached and the potential field signals "smear" (e.g., gravity decreasing towards the granitiod body). The relationship observed between RF uncertainty and distance to geological boundaries is consistent with prior observations (e.g. Cracknell et al., 2013). High H is also observed in the southeast region of the study area. It is not possible to determine whether the interpretation map is incorrect, however, both the RF classifications and high H suggest that this rock unit is significantly more complex than shown. This is a clear example of the benefit of the analysis of RF classification in conjunction with uncertainty and may serve to optimise ongoing field efforts, either outcrop mapping or drilling as appropriate.

3.6 Conclusions

This study demonstrates that Random Forests (RF) may be applied to reconnaissance type geophysical data, in the absence of geochemistry, and produce sound lithological predictions. There are two obvious applications for the use of RF for early stage geological mapping. The first is for the refinement of an existing geological map. The second being the production of a geological map from a limited number of observations in the creation of a first pass map. Sparse outcrop or a broad drilling campaign could provide such starting observations, provided the spatial distribution of observations adequately samples the project area.

In this demonstration study, RF was able to preserve class labels i.e. stratigraphic context where more than one class comprised the same lithology. This is an important outcome as the timing relationships between mineralisation and various stratigraphy are vital information for mineral prospecting. Proterozoic dykes, which are petrophysically indistinguishable from Archean mafic rocks in the study area, confusing aeromagnetic interpretation. RF by utilising a higher dimensional data space can deal with this complication, provided examples of the dykes overprinting the older stratigraphy are sampled in the training data.

H provides a valuable insight into classification results. The highest H denotes areas of geological/geometric complexity and proximity to lithological boundaries. Where a predicted lithological boundary significantly differs from the reference map, the behaviour of H proximal to interpreted and predicted boundaries indicates which position is most probable. Statistically distinct populations in H correlate with correctly and incorrectly classified samples. Through understanding H, an optimal trade-off, retaining the greatest number of correct samples whilst discarding incorrect samples can be identified. Understanding the distribution of H for correct and incorrect sample populations allows a user to define an acceptable

trade-off between discarding the maximal number of incorrectly classified samples or retaining a more complete, albeit potentially less accurate map. This will reflect the tolerance for risk of each individual explorer/company. The combination of RF classification and uncertainty appraisal allows explorers to critique quantitatively, the validity of map outputs: a quality control measure not available in conventional mapping.

3.7 Acknowledgements

We would like to thank Gold Fields Ltd. for access to data for the purpose of this study. Stephen Kuhn is supported by an Australian Postgraduate Award Scholarship from the University of Tasmania. This research was conducted in collaboration with the ARC Industrial Transformation Research Hub for Transforming the Mining Value Chain (project number IH130200004) at the Centre of Excellence in Ore Deposits, University of Tasmania. The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council. We utilised the Orange software package (Demsar et al., 2013) for RF classification. Pre-processing, interpolation and plotting were performed using Geosoft Oasis Montaj and ESRI ArcGIS. David Doutch is thanked for his input on the geological and structural setting of the project. We thank the Assistant Editor, Associate Editor and 3 reviewers for their suggestions which have significantly improved the manuscript.

References

Barnett, C. and Williams, P. (2009). Using geochemistry and neural networks to map geology under glacial cover, *Geoscience BC Report*. accessed 15 November 2017. **URL:** *http://www.geosciencebc.com/i/project*$_d$*ata/QUESTdata/GBCReport*2009 − 3/*GBC_Report_*2009 − 3.*pdf*

Blewett, R., Squire, R., Miller, J., Henson, P. and Champion, D. (2010). Architecture and geodynamic evolution of the St Ives Goldfield, eastern Yilgarn Craton, **183**: 275–291.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

Breiman, L. (2001). Random Forests, *Machine Learning* **45**: 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth Brooks/Cole Statistics/Probability Series, Wadsworth International Group.

Connors, K., Stolz, E. and Hanneson, J. (2002). Early fault architecture at St Ives: implications for Au and Ni mineralisation, *Applied Structural Geology for Mineral Exploration and Mining: Australian Institute of Geoscientists, Bulletin* **36**: 29–31.

Cox, S. and Ruming, K. (2004). The St Ives mesothermal gold system, Western Australia — a case of golden aftershocks?, *Journal of Structural Geology* **26**: 1109–1125.

Cracknell, M. (2014). Machine learning for geological mapping: Algorithms and applications.

Cracknell, M. and Reading, A. (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): 113 – 126.

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

Cracknell, M., Reading, A. and McNeill, A. (2014). Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests^{TM} and Self-Organising Maps, *Australian Journal of Earth Sciences* **61**: 287–304.

Crawford, M. (2011). Dynamic coupling between deformation processes, fluid-rock interaction, and gold deposition in the Argo gold deposit, St Ives.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. and Zupan, B. (2013). Orange: Data mining toolbox in python, *Journal of Machine Learning Research* **14**: 2349–2353.

Grunsky, E. and Kjarsgaard, B. (2016). Recognizing and validating structural processes in geochemical data, *Compositional Data Analysis: Springer Proceedings in Mathematics and Statistics* p. 85–116.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics, Springer.

Kuhn, S., Cracknell, M. and Reading, A. (2008). Lithological mapping via Random Forests: Information entropy as a proxy for inaccuracy, *25th International Geophysical Conference and Exhibition, ASEG, Extended Abstracts* p. 1–4.

Miller, J., Blewett, R., Tunjic, J. and Connors, K. (2010). The role of early formed structures on the development of the world class St Ives Goldfield, Yilgarn, WA, *Precambrian Research* **183**: 292–315.

Ngyuen, T. (1997). Structural controls on gold mineralisation at the Revenge mine and its tectonic setting in the Lake Lefroy area.

Rodriguez-Galiano, V., Chica-Olmo, M. and Chica-Rivas, M. (2014). Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, southern Spain, *International Journal of Geographical Information Science* **28**(7): 1336–1354.

Ruming, K. (2006). Controls on lode gold mineralisation in the Victory thrust complex, St Ives Goldfield, Western Australia.

Shannon, C. (1948). A mathematical theory of communication, *Bell Systems Technical Journal* **27**: 379–423.

Squire, R., Allen, C., Cas, R., Campbell, I., Blewett, R. and Nemchin, A. (2010). Two cycles of voluminous pyroclastic volcanism and sedimentation related to episodic granite emplacement during the late Archean: Eastern Yilgarn Craton, Western Australia, *Precambrian Research* **183**: 251–274.

Swager, C. (1997). Tectono-stratigraphy of late Archaean greenstone terranes in the southern Eastern Goldfields, Western Australia, *Precambrian Research* **83**: 11–42.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer-Verlag, New York.

Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley Sons Inc.

Waske, B., Benediktsson, J., Árnason, K. and Sveinsson, J. (2009). Mapping of hyperspectral aviris data using machine-learning algorithms, *Canadian Journal of Remote Sensing* **35**: 106–116.

Wellmann, J. and Regenauer-Lieb, K. (2012). Uncertainties have a meaning: Information entropy as a quality measure for 3D geological models, *Tectonophysics* **526–529**: 207–216.

Yu, L., Porwal, A., Holden, E. and Dentith, M. (2012). Towards automatic lithological classification from remote sensing data using support vector machines, *Computers and Geosciences* **45**: 229–239.

Chapter 4

Identification of intrusive lithologies in volcanic terrains in British Columbia by machine learning using Random Forests: The value of using a soft classifier

Stephen Kuhn[12], Matthew J. Cracknell[12], Anya M. Reading[23] and Stephanie Sykora[4]

[1] Centre for Ore Deposit and Earth Sciences (CODES) University of Tasmania.
[2] ARC Industrial Transformation Research Hub for Transforming the Mining Value Chain (TMVC), University of Tasmania.
[3] School of Natural Sciences (Maths and Physics), University of Tasmania.
[4] First Quantum Minerals Ltd.

4.1 Abstract

Identifying the location of intrusions is a key component in exploration for porphyry Cu Mo Au deposits. In typical porphyry terrains, in absence of outcrop, intrusions can be difficult to discriminate from the compositionally similar volcanic and volcanoclastic sedimentary rocks in which they are emplaced. The ability to produce lithological maps at an early exploration stage can significantly reduce costs by assisting in planning and prioritisation of detailed mapping and sampling. Additionally, a data-driven strategy provides opportunity for the discovery of intrusions not identified during conventional mapping and interpretation. We used Random Forests, a supervised machine learning algorithm, to classify rock types throughout the Kliyul porphyry prospect in British Columbia, Canada. Rock types determined at geochemical sampling sites, were used as training data. Airborne magnetic and radiometric data, geochemistry and topographic data were used in classification.

Results were validated using First Quantum Minerals' geological map, which includes additional detail from targeted location and transect mapping. The petrophysical and compositional similarity of rock types resulted in a noisy classification. Intrusions, particularly the more discrete, were inconsistently predicted, likely due to their limited extent relative to data sampling intervals. Closer examination of class membership probabilities identified locations where the probability of an intrusion being present was elevated significantly above

background. Indeed, a large proportion of mapped intrusions correspond to areas of elevated probability and importantly, areas were highlighted as potential intrusions which were not identified in geological mapping. The Random Forests Classification produced a reasonable lithological map, if lacking in resolution, but more significantly, great benefit comes from the insights drawn from the Random Forests class membership probabilities. Mapping the spatial distribution of elevated intrusion class membership probability, a soft classifier approach, produced a map product which can target intrusions and prioritise detailed mapping for mineral exploration.

4.2 Introduction

4.2.1 Regional geology

Northern British Columbia, Canada, dominantly consists of a series of intra-oceanic island arc terrains that were accreted onto ancestral North America in the Mesozoic (e.g., Bond and Kominz, 1984; Gabrielse et al., 1991; Monger and Price, 2002; Nelson and Colpron, 2007; Johnston, 2008). The Kliyul Cu - Au prospect is located in the northern end of one such terrain, known as Quesnellia (Figure 4.1). Quesnellia consists of high-potassium calc-alkaline to shoshonitic submarine volcanics and sediments known as the Takla Group (Lord, 1948; Monger, 1977). The Hazelton Group andesitic volcanics, which are abundant at the Kemess porphyry Cu - Au deposit north of Kliyul (Gordee et al., 2004; Rebagliati et al., 1995) unconformably overly the Talka group. The Hogen Batholith is located to the south of Kliyul, and is a large, fertile, late Triassic to Early Cretaceous igneous body that hosts other porphyry deposits such as the Lorraine (Woodsworth, 1976; Nelson and Bellefontaine, 1996).

4.2.2 Local geology

A new geological map of the Kliyul property (Figure 4.2) was created following a 350 m gridded rock sample campaign of the entire property, combined with targeted 1:10,000 scale mapping during August 2017 by First Quantum Minerals and AuRico Metals personnel. This map, showing features as described below, will constitute the reference geology map with which Random Forests (RF) outputs will be compared. The Kilyul property consists of volcano-sedimentary strata, a series of discrete to large intrusive bodies and several prominent faults (Table 4.1; Figure 4.2). Quaternary glacial-fluvial cover fills the low-lying valleys. Volcano-sedimentary strata at Kliyul consist of two units of the Takla Group volcanics, Goldway Peak (GP) and Kliyul Creek (KC) Groups.

These are equivalent to the same named units in the British Columbia Geological Survey government map of the area (Schiarizza, 2004a, b; Schiarizza and Tan, 2005). The oldest unit is KC which consists of four sub-units: intermediate to mafic volcanics and volcanoclastics (KCv); carbonate-rich sediments and volcanics (KCc); felsic volcanics and volcanoclastics interlayered with carbonate-rich sediments (KCfc); and siliceous sediments (KCs). Strata have dominantly low-angle north-east dips.Overlying KC and all its sub-units are the GP volcanics. GP consists of more mafic to intermediate augite-phyric volcanic flows and breccias. A series of calc-alkaline, fine-to-coarse grained intrusions with variable pre-, syn-to-post mineralization and alteration relationships cross-cut the KC and GP

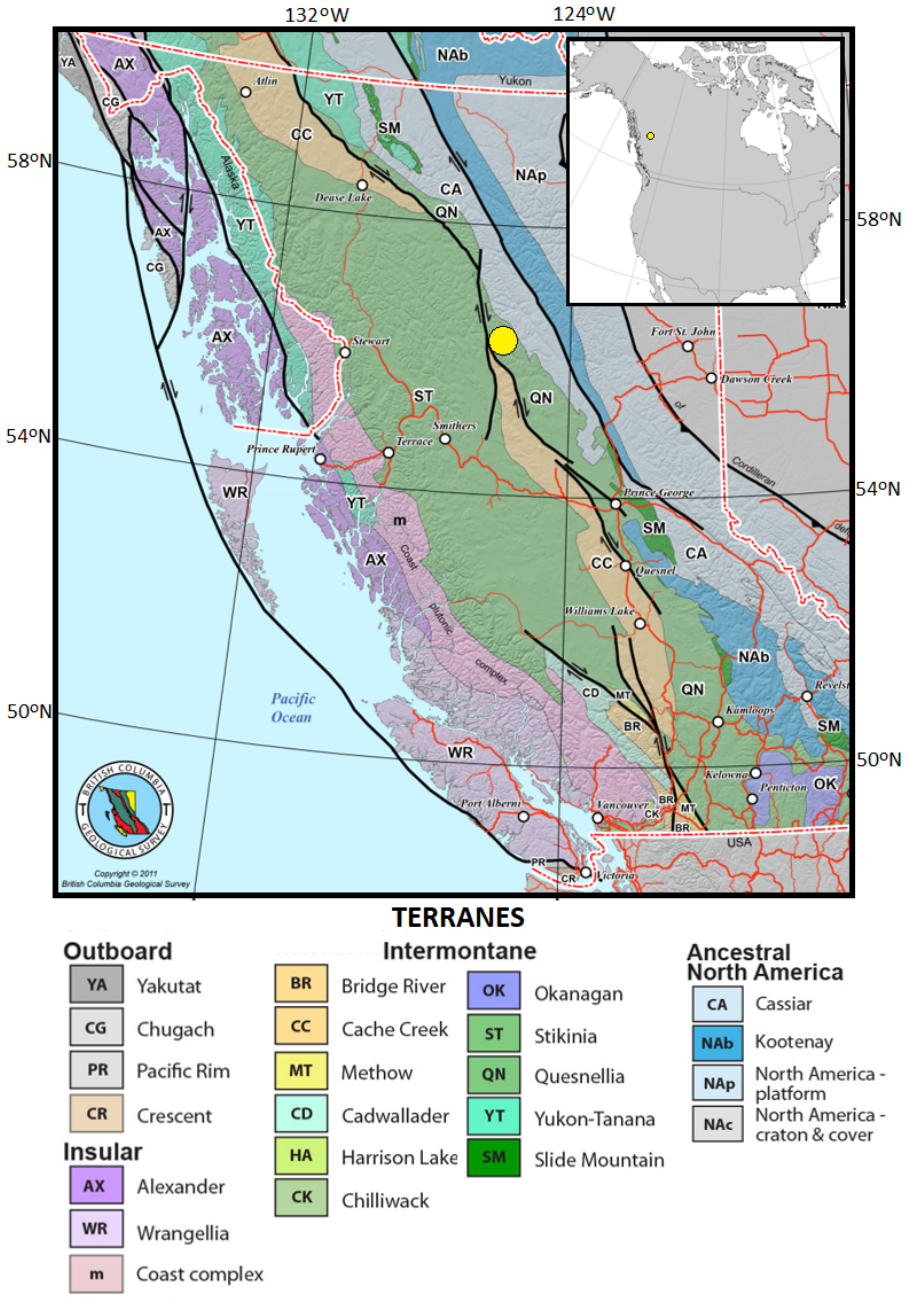


FIGURE 4.1: regional geology of British Columbia, Canada (modified from Nelson and Colpron, 2011). the location of the Kliyul porphyry Cu – Au prospect is within the yellow marker.

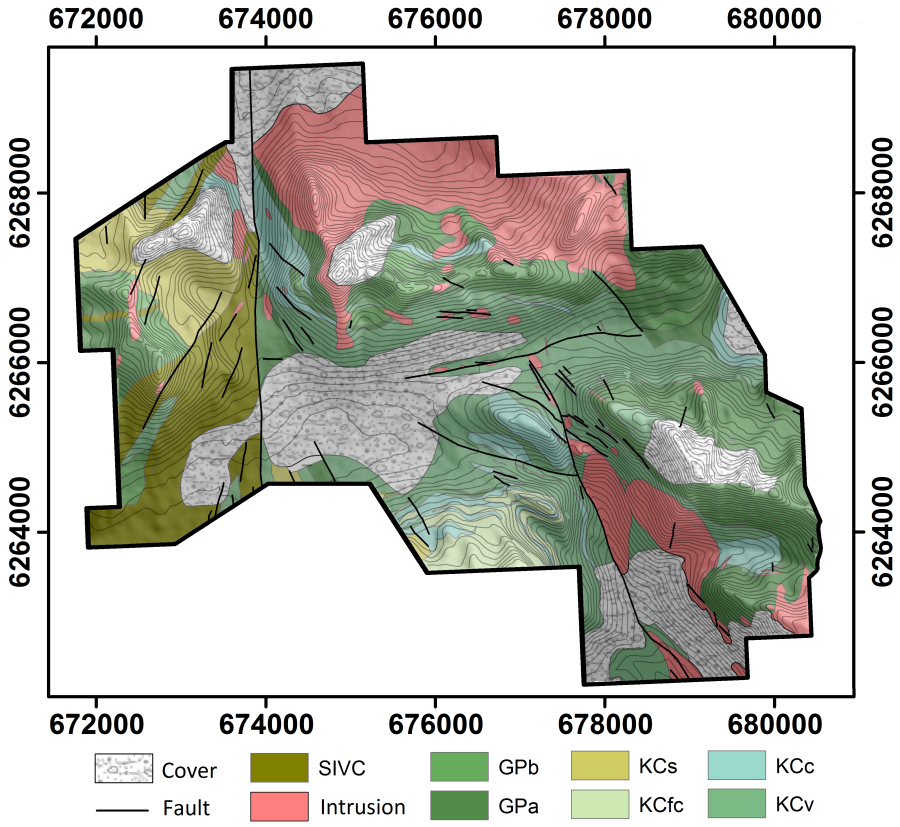


FIGURE 4.2: Geology map of the Kliyul project, draped on topography (SRTM, shown with 20m contour). Lithology codes are detailed in Table 4.1.

volcano-sedimentary strata (Figure 4.2). Age dating is poorly constrained, but historic and recent efforts estimate a Late Triassic to Early Jurassic timing (217 Ma). Three apparently long-lived structures influence the map patterns at Kliyul: (1) the north-northwest striking Kliyul Creek Fault with an apparent dextral sense of shear with both sinistral and dextral fault splays; (2) the east-northeast-striking Valley Fault with poorly constrained, but apparent normal, north-block-down kinematics, and; (3) the north-striking dextral Dortatell fault which had a strong asymmetrical influence of rock types to the west versus east of the fault, with a mylonitic-like shearing of the rocks in the southwest.

4.2.3 Random Forests

We use Random Forests (RF; Breiman, 2001), a supervised machine learning algorithm, to classify each pixel in the study area according to lithology. A number of algorithms have been tested and applied to geological mapping enterprises. Examples include variants of artificial neural networks (e.g. Barnett and Williams, 2009; Cracknell et al. 2014; Grunsky and Kjarsgaard, 2016), support vector machines (Yu et al., 2012; Cracknell et al., 2014) and Random Forests (Cracknell and Reading, 2013; 2014; Cracknell et al., 2014; Rodriguez-Galiano et al., 2014; Harris and Grunsky, 2015; Kuhn et al., 2016; 2018). While all of these choices have been successfully deployed in various scenarios, RF has been consistently identified as a good choice of classifier on grounds of both performance, ease of use (both in terms of few and simple tuning parameters and insensitivity to variable scaling) and

TABLE 4.1: Simplified stratigraphy of the Kliyul project. Lithology and colour codes shown in this table will be used for all figures in this study.

Group	Unit	Unit Code
Cover	Ice cover / Overburden	Cover
SIVC	Mixed zone of intense shearing	SIVC
Intrusion	Intrusions (various)	Intrusion
Takla Group	Goldway Peak	GPb
		Gpa
	Kliyul Creek - Siliceous sediment subunit	KCs
	Kliyul Creek - Felsic volcanic bearing subunit of the KCc	KCfc
	Kliyul Creek - Carbonate-rich subunit	KCc
	Kliyul Creek - Volcanics and volcanicalstics	KCv

ability to handle data with high intra-class variability and high interclass similarities (Cracknell and Reading, 2013; Breiman, 2001) by multiple studies. These include using RF alone (e.g. Cracknell et al., 2014; Kuhn et al., 2016, 2018) or in comparison with other algorithms (Cracknell and Reading, 2014; Waske et al, 2009). The accessibility of RF makes the methods described in this study widely applicable in the geoscientific disciplines, minimising the barrier to entry imposed by requirements for specialised computing skills and equipment.

RF extends upon the classification and regression tree (CART) methodology (Breiman et al., 1984), addressing the high variance (Friedman, 1997) that can be associated with single decision trees (Murphy, 1998) by constructing an ensemble or forest of pseudo-unique decision trees (Figure 4.3). The performance of a RF is controlled by the strength of individual trees in the forest and the correlation between trees, with a forest of strong, de-correlated trees being the ideal case (Breiman, 2001; Hastie et al., 2009).

To minimise correlation between trees, randomness is introduced into the selection of subsets shown to each tree and in the selection of variables used to split decision nodes. Bagging (Breiman, 1996) is used to produce a random subset of input training data, of equal size to the training set provided to each classification tree. This has been shown, on average, to include approximately 63.7% of unique instances from the training set, with the remainder, 'out of bag' instances, held aside for testing (Breiman, 2001). During learning, a randomly selected pool of variables of pre-determined size is provided to each decision node. From this pool, the variable which produces the maximal improvement in homogeneity of the child nodes relative to the parent node, as measured by Gini impurity (Breiman et al., 1984 and described by Kuhn et al., 2018), is used to split that node (Figure 4.4.3). This split is performed at all nodes in each classification tree and subsequently, all trees in the Random Forest. Final classifications are assigned as the mode of the forest (or mean the regression case).

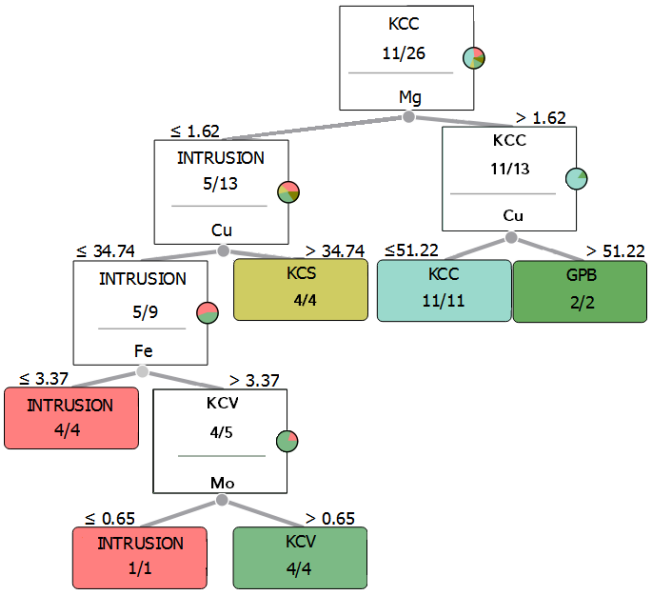


FIGURE 4.3: Schematic representation of a small end segment of one of the 500 classification trees comprising the RF trained in this study. Each node shows, from top to bottom: the modal class, the number of samples comprising the modal class, the variable used to subdivide the node and the value at which that subdivision occurred. Nodes are coloured by lithology (as given in Figure 4.2 and Table 4.1) when homogeneity is reached.

4.2.4 Class membership probabilities and uncertainty

RF acts as a soft classifier: the final decision is determined as the majority vote of many competing solutions. The decision of individual trees and the proportion trees voting for each class: class membership probabilities (Breiman, 2001; Hastie et al., 2009), are recorded. Class membership probabilities, in addition to being fundamental to the operation of RF, represent valuable metadata in assessing the strength and distribution of classification results. A class may reach majority by a narrow margin, highlighting the importance assessing class membership probabilities. It is possible that a class other than the majority solution, could have exhibited a near equal probability. The spread of class membership probabilities can be made use of to quantify uncertainty. In this study, we calculate information entropy (H; Shannon, 1948), as defined by:

$$H = -k \sum_{i=1}^{n} p_i log p_i$$

where p_i is the class membership probability at location i, n is the number of candidate classes, k is a positive constant. Both k and the logarithm base are arbitrary scaling constants. H is a measure of disorder in a system with minimal H corresponding to complete homogeneity and maximal H corresponding to a heterogeneous spread with each class equally represented. H is monotonic, increasing with the number candidate classes added. It has been demonstrated that H is a useful metric in assessing the spatial distribution of uncertainty in lithological classification made using RF (Kuhn, et al., 2016) in addition to prior usage in displaying the per-voxel uncertainty associated with 3D potential field inverse models (Wellmann and Regenauer-Lieb, 2012). Kuhn et al (2016) used a

normalised form of H (H_{norm}), accounting for the number of classes possible for a given sample, representing how closely each sample approaches their own respective minimal or maximal state of disorder. As such, a sample in which all classes are equally probable, will have an H_{norm} equal to 1, regardless of how many classes were present. Using these metrics, we can describe both the complexity of a system (H) and make an inference about the possible accuracy of the classification (H_{norm}), as demonstrated by Kuhn et al. (2016). Assessment of classification validity is made with reference to the First Quantum Minerals 1:10000 scale geological map (henceforth referred to as FQM geological map) until further ground truthing can occur. What is, in effect, being measured is the consistency between maps. In such a case, either the RF classification or the FQM geological map (or both) could be in error. Cracknell and Reading (2014) and Kuhn et al. (2016) provide examples of how the behaviour of H and H_{norm} may aid in such an assessment.

4.2.5 Objectives

In this study we explore a pragmatic approach for a relatively data-limited situation. We train a RF classifier using soil sample points defined by geochemical and geophysical data. This RF is subsequently used to classify all remaining samples in the Kliyul project area. The soft classification metric, class membership probability, and the more usual classification results are used to predict lithology and identify intrusions in the project area. H is used to define, uncertainty and complexity associated with classification objectively. We employ the FQM geological map, which benefits from rock and soil sampling, in addition to detailed location and transect mapping, as a benchmark with which to compare classification results.

4.3 Methods

4.3.1 Data and sampling

This study incorporates all available geochemical and geophysical datasets that encompass the project in addition to shuttle radar topography mission (SRTM) elevation data. Geochemical data includes a suite of 49 measured elements (ICP21 + MS61 + XRF5000) for rock and soil samples collected at a 350 m x 350 m spacing over the project. These have been pre-processed such that all data exhibiting a measurement below detection limit were assigned a value of half of the detection limit for that element. Datasets comprising an excessive number of samples below detection limit (in excess of 25 percent or where spatial distribution resulted in large areas with insufficient real data) were omitted.

Geophysical data (Supplement 1) comprised reduced to pole (RTP) total magnetic intensity data and derived datasets (vertical and horizontal derivatives, tilt derivative, total horizontal derivative, analytic signal and analytic signal of the vertical integral), and airborne radiometric data (Potassium, Thorium, Uranium and total count), each collected at a 100m line spacing and gridded at 25m. IP/resistivity and electromagnetic data have been collected in this area however their extent was limited, lacking coverage over the entirety of the study area and thus they were not included in this exercise (the present study could be used to target future surveys of this kind with considerable cost savings).

All datasets were resampled to a 50 m x 50 m grid and compiled as a matrix in the form of easting, northing, V_1, V_2, ..., V_n (where V is a given variable/dataset). In order to preserve conditional independence, where pairs of variables correlated in excess of 0.8 (Pearson's correlation coefficient), one of the pair was removed. Where a variable was a dataset exhibiting any indication of poor quality, e.g. excessive missing data or readings below detection limit, that dataset was omitted. Where multiple correlations where present, datasets were removed in such a way as to maximally reduce the dimensionality of the dataset. If neither of these criteria were encountered, but only where a correlation in excess of 0.8 was shown, datasets were removed based on subjective geological utility. The datasets analysed as important for best possible results (as described in the following section) are shown in Table 4.2.

TABLE 4.2: Datasets used in this study, ranked in order of importance (as indicated by rank and corresponding score) by Random Forests. 10-fold Accuracy (scaled from 0 to 1) describes the 10 fold cross validation accuracy achieved by Random Forests when including a given variable in addition to all those ranked higher. For example, when using an RF trained using variables 1 to 10, as was used in this study, a 10 fold cross validation accuracy of 0.835 is achieved.

Rank	Dataset	Score (RF)	10-Fold Accuracy
1	K	46	0.632
2	Fe	44	0.692
3	Ca	43	0.73
4	Mo	41	0.762
5	Rb	36	0.791
6	RTP	32	0.8
7	Mg	32	0.807
8	Zr	32	0.816
9	Pb	30	0.824
10	**Ba**	**29**	**0.835**
11	Cu	28	0.828
12	Ga	28	0.836
13	Ag	27	0.831
14	W	26	0.837
15	Mn	26	0.834

Pixels comprising samples, for which a lithology had been assigned (Table 4.1) were taken as training data. The remaining data were held for classification by the trained RF. RF is prone to bias in favour of a numerically dominant class (Hastie et al, 2009). In order to mitigate this tendency, classes were balanced to 50 samples per class; through either bootstrapping (where less than 50 samples were available) or random decimation (where more than 50 samples were available.

The composition of intrusions is highly variable and comprise both felsic, intermediate and mafic units. Due however to the impractically small number of samples for each, this class was combined. We confirm that the training data adequately samples the intrusion class and that training data samples and intrusion

types are co-located. Data partitioning considerations in the spatial context, in a general sense applicable to machine learning, are outlined by Vucetic et al. (1999).

The main objective of this study is to determine if RF can identify intrusions from data typical of early-stage exploration, despite the heterogeneous class definition (and also noting that we aim to identify an indicator lithology at this early stage, not the ore deposit itself). It is understood that better, and worse, results could be obtained through other means of retaining or adding synthetic samples to achieve a balanced class size. For the purpose of the present study, this process was restricted to the balancing described above, similar to the method used by Kuhn et al. (2019) using only observed data values as opposed to imputing from an estimation of a datasets probability density function or other common methods. In this study, at an early stage in the exploration cycle, the notable limiting factor is the resolution of the input data, however, it would be expected that results would improve were additional training data available (Cracknell et al., 2013). Samples comprising the training set are shown in Figure 4.4.

4.3.2 Variable ranking, reduction, and definition of Random Forest classifier

Following the balanced sampling process, RF is used to rank the importance of variables according to the mechanism described by Breiman (2001). Under this strategy, each variable is permuted and shown to the RF (Demsar et al., 2013). That variable which produces the largest difference in classification accuracy is most important and the variable which, when permuted, produces the smallest difference is least so. In this study, we seek to produce a classifier which benefits from the additional dimensionality permitted by a machine learner as compared to more conventional methods (for example, successive comparison of scatter plots or a manually weighted GIS analyses) while producing a result that maximises interpretability by the end user. As such, we seek to reduce the number of input variables to the minimum required to produce the best result. This is possible as RF tends to produce peak accuracy with the inclusion of a given number of variables, after which, results remain stable or may even marginally deteriorate; additional variables at this point are redundant. Prior studies deploying RF in a similar fashion for lithology classification (e.g. Kuhn et al, 2019, Kuhn et al., 2018; Kuhn et al. 2016; Cracknell et al., 2014) have shown a tendency for this to occur at between 8 and 15 variables though this may change as additional studies are performed.

For this study, we used a RF comprising 500 classification trees with no pruning or growth restrictions (see also Supplement 2 for other implementation parameters). RF is not prone to overfitting with additional trees, instead reaching a stable error minima (Breiman, 2001). This number of trees is well above what is likely required and represents a safe choice for geoscientists looking to replicate the methodology of this case study without any risk of using insufficient trees. These parameters were duplicated during ranking. All variables were ranked via the process described prior. These variables were then successively used to build a classifier which was in turn assessed for accuracy. We used 10-fold cross validation (James et al., 2013) in conjunction with a backwards recursive process to determine the cross-validation accuracy of the RF, drop off the lowest ranked variable and re-rank variables. This procedure was repeated for all variables. Best results of 83.5% cross validation accuracy were achieved with the inclusion of the best ranked 15

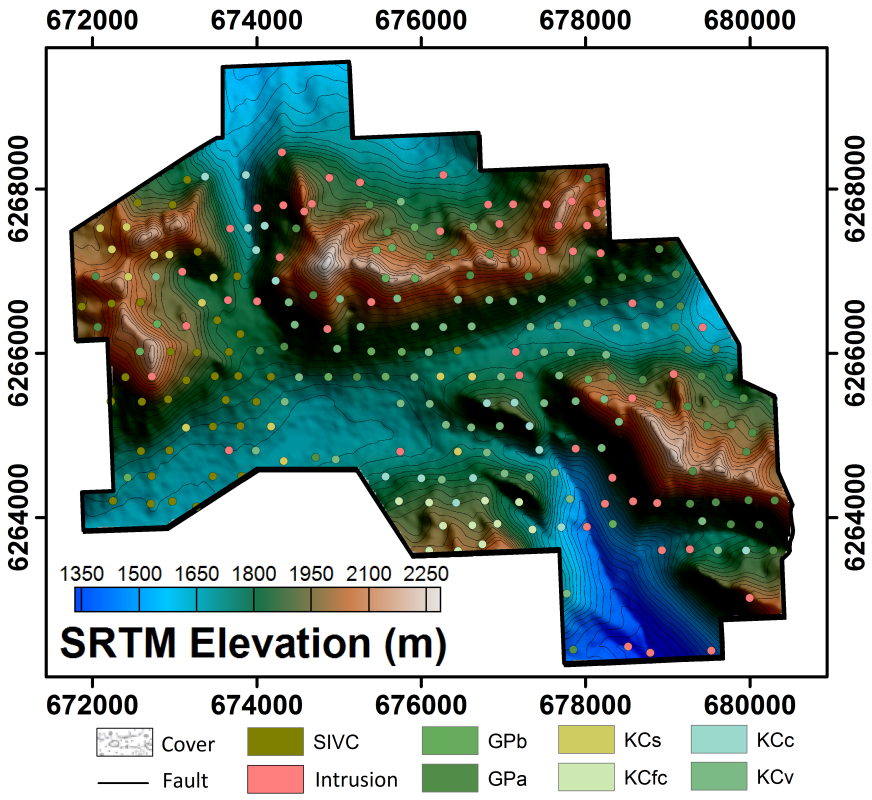


FIGURE 4.4: Training sample locations coloured by lithology underlain by the SRTM DTM (shown with a 50m contour). Lithology class names are given in Table 4.1.

variables (Table 4.2). Examples of top ranked geophysical and geochemical datasets are shown in Figure 4.5.

Using a confusion matrix (Table 4.3; a comparison of actual class labels with predicted class labels for each training data point is made, in this case during 10 fold cross validation), we can see that several classes were consistently classified correctly. The KCv class performed the most poorly, being commonly misclassified as intrusive or GPa. Intrusions were most often misclassified as KCv. The RF used to produce this result was selected as the final classifier for use on all remaining data. While the relationships shown in this confusion matrix (based on cross validation results) do not necessarily translate to classification on blind data, this does give some indication of the potential strengths and weaknesses of the selected classifier.

4.4 Results

The results of the RF classification show 73% overall consistency (Figure 4.6A) with the FQM geological map (Figure 4.2). Results were more consistent with the FQM geological map where a lithology was present in larger domains, whereas results appear less robust for narrower zones of a given class, such as where lithologies appear to wrap around the topography due to low angle bedding dips. This is in part a function of the resolution of the geochemical input data and is discussed in the next section. H identifies many of the class boundaries in the project area while

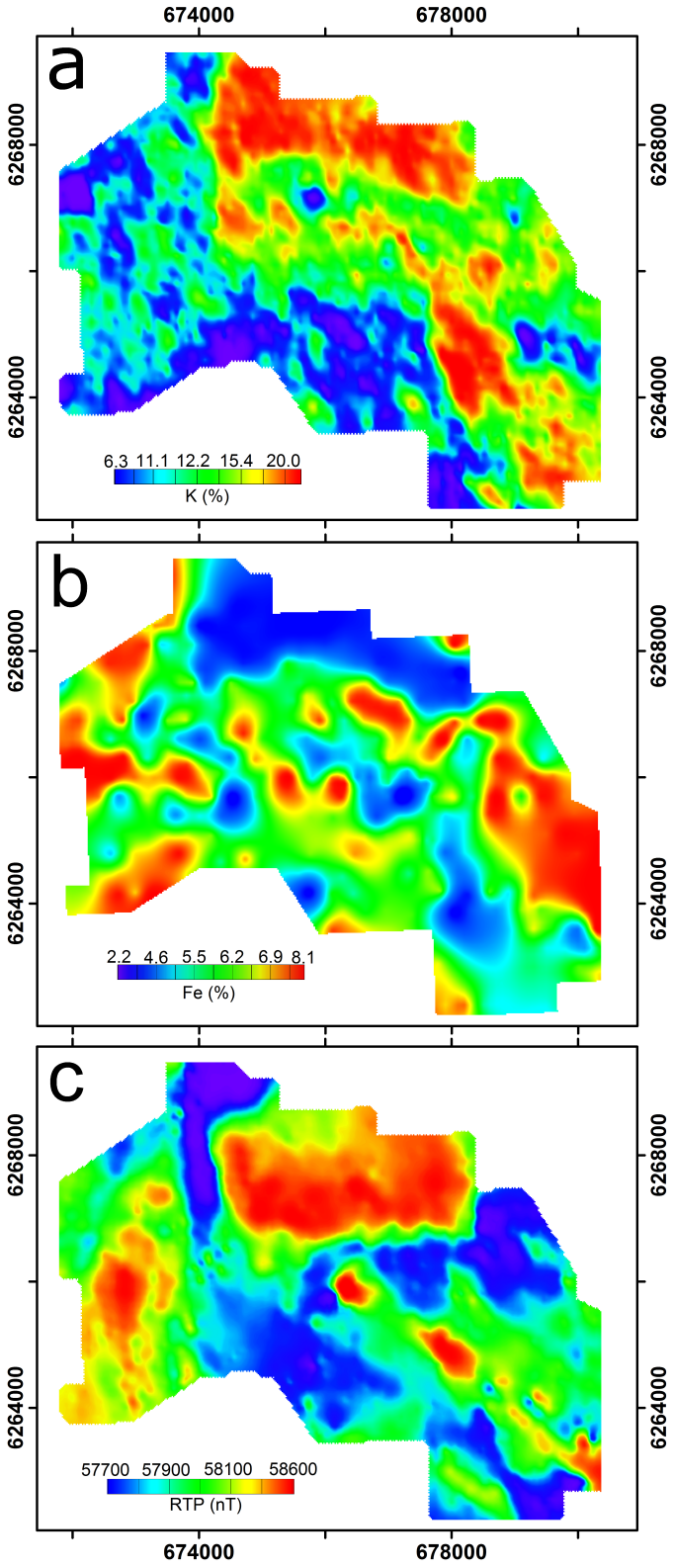


FIGURE 4.5: Example of datasets used in this study: A) K (radiometric, % K), B) Fe (% Fe in sample), C) Reduced to pole total magnetic intensity.

TABLE 4.3: Confusion matrix showing the performance of the Random Forest classifier used in study on the provided training data. This is useful in assessing classifier performance and drawing inference about class similarity/dissimilarity and where misclassification is likely to occur but is not a measure of performance on new data.

		PREDICTED								
		Gpa	GPb	Intrusion	KCc	KCfc	KCs	KCv	SIVC	Total
OBSERVED	GPa	**33**	4	4	0	1	2	2	4	**50**
	GPb	0	**50**	0	0	0	0	0	0	**50**
	Intrusion	3	0	**34**	0	1	1	9	2	**50**
	KCc	0	0	0	**50**	0	0	0	0	**50**
	KCfc	0	0	0	0	**50**	0	0	0	**50**
	KCs	0	0	0	0	0	**50**	0	0	**50**
	KCv	6	0	6	2	4	1	**27**	4	**50**
	SIVC	3	1	2	2	0	2	2	**38**	**50**
	Total	**45**	**55**	**46**	**54**	**56**	**56**	**40**	**48**	**400**

normalised information entropy (H_{norm}) shows a high per-pixel uncertainty throughout the project area (Figure 4.6B, 4.6C). Both variants highlight some areas for which RF classified with low uncertainty, including, for example, the large intrusion in the north of the project. Class membership probabilities provide a more detailed description of the information summarised in the calculation of H and H_{norm}. It can be seen (Figure 4.7) that there are clear domains classified as Intrusive in the north, GPa in the east, SIVC in the west and KCv through the centre of the project. It can be seen, however, that as demonstrated by H_{norm} (Figure 4.6), multiple classes were similarly probable particularly in the centre of the project area. It is worth noting that many discrete zones of elevated probability of Intrusive class, the target of economic significance, are identified. In many of such cases, these areas were not classified in the final RF lithology map as intrusive. The value of using the soft classifier, elevated class membership probability, is clear from this analysis.

4.5 Discussion

Many of the broader domains in the RF lithology map resulting from the RF classification are consistent with the FQM geological map. The most notable shortcomings in the RF lithology map are due to a lack of resolution in regions where the mapped lithotypes have more detail than the geochemical sample spacing which is reflected in the pixel size used in this study. Large intrusions were identified in final classification, therefore many of the subtler intrusive features were either not identified or mapped without adequate resolution to be of substantial use in targeting or follow up work. There are several likely causes for this, foremost of which is the resolution of input data, a result of the sampling interval of each dataset and the effects of interpolation of those data.

The limitations of gridding geochemical data are well understood as a potential source of error in the result. If attempting to predict the nature of a region not represented by a sample, a value must be imputed or interpolated. Interpolation of

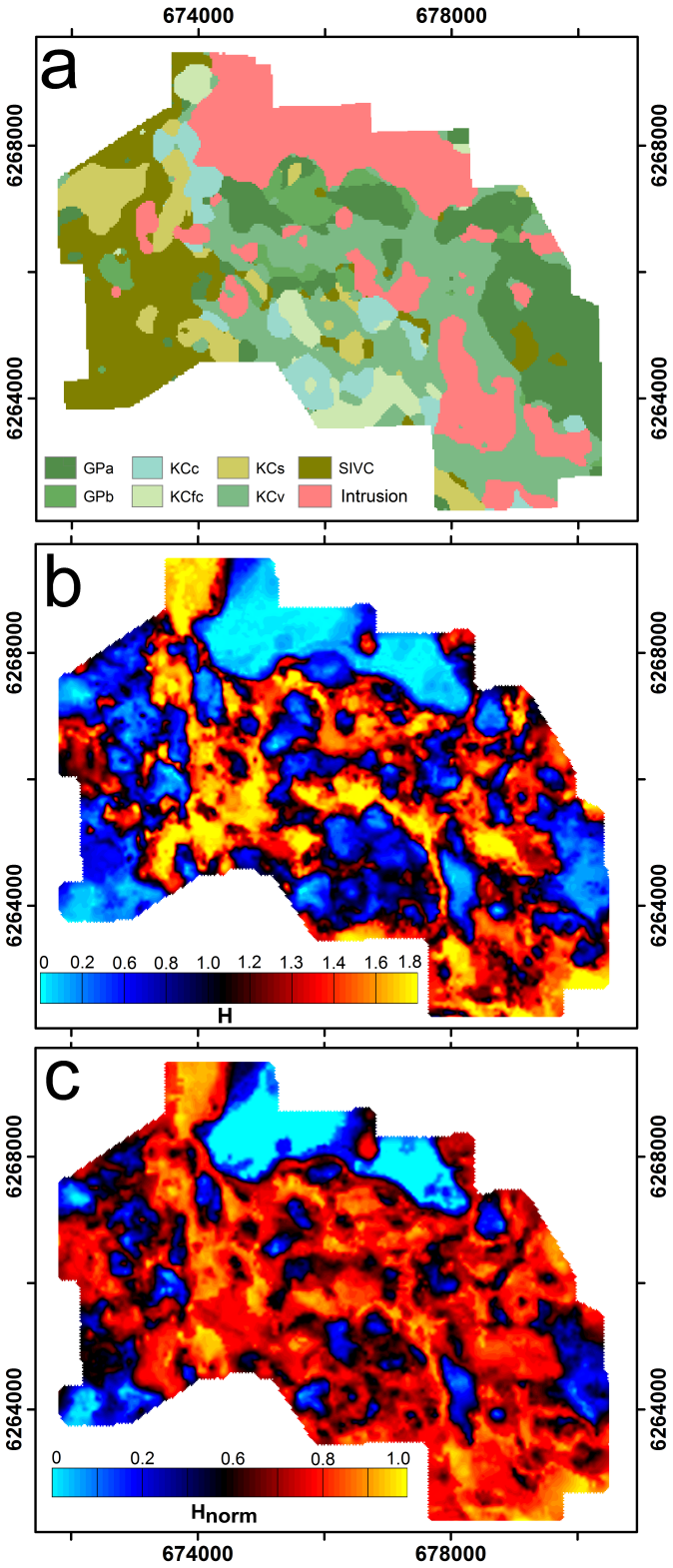


FIGURE 4.6: A) RF classification, shown smoothed by a modal convolution filter (3x3 kernel), B) information entropy (H), C) information entropy normalised by number of classes per-pixel (H_{norm}).

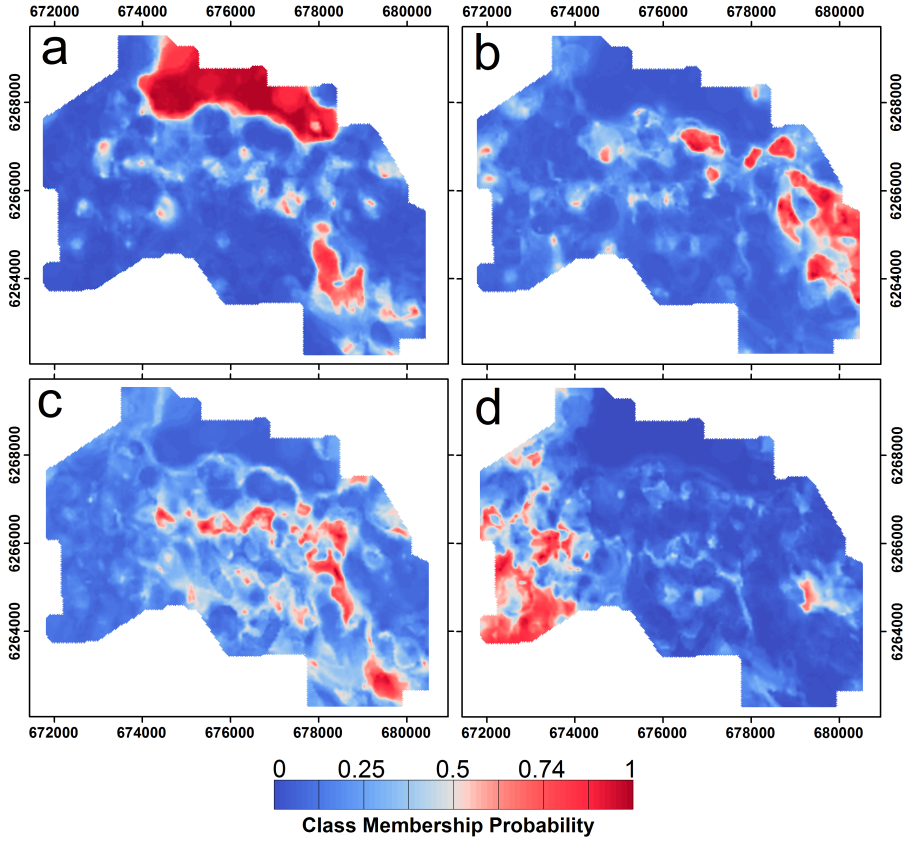


FIGURE 4.7: The relative proportion of trees in the Random Forest voting for each class (class membership probabilities). Examples shown for: A) Intrusions, B) GPa, C) KCv and D) SIVC.

geochemical data is a technique used pervasively throughout the mineral exploration industry. As such, this study provides a faithful indication of performance on the data likely to encountered in a typical industry case example. More conventional means of interpreting magnetic data (e.g. Isles and Rankin, 2013; Salem et al., 2008) could potentially be used to further improve RF interpretation. While in this case all magnetic derivatives were discarded due to poor ranking on ability to define lithology, these could be used to better define structure and boundaries throughout the project area. It is entirely plausible that the use of more sophisticated imputation strategies, in place of simple bootstrapping could improve training class definition and thus results, particularly where available real samples facilitate the accurate prediction of a given classes' true distribution. This principle extends to several aspects of the study methodology. This study was intentionally designed to demonstrate a rapid, effective approach easily adopted by the wider geoscientific community, and produced good results. It is possible that further tuning of RF parameters, or more advanced feature ranking and selection (e.g. Lundberg and Lee, 2017) might be useful for studies in a more data-rich context.

The use of an ensemble of unique classifiers (classification trees) is the key feature of RF. Not only does this have the important benefits of improved accuracy and the ability to respond well in dealing with high intraclass variability and interclass similarity; but this also facilitates a detailed analysis of classification results.

Plotting individual class membership probabilities indicates where each class was most likely to occur (Figure 4.7), and in this study is an approach that leads to valuable insight for the exploration program going forward. H responds both to uncertainty and complexity: the distribution of probabilities and the number of classes possible at a given sample. In this case, observing H (Figure 4.6B) highlights boundaries between classes and other regions of more complex geology. This coincides with similar observations of the behaviour of H (e.g. Kuhn et al., 2016, 2019). H_{norm} (Figure 4.6C), a measure of how closely each instance approaches its own maximal uncertainty, is high throughout much of the study area. This suggests that there is a high likelihood of the classification being incorrect (e.g. Kuhn et al., 2016, 2018).

Given the stated goal of identifying and defining intrusions, potentially of economic significance to the project, we focused on the class membership probability (CMP) of this unit (Figure 4.7). Even where this is not the majority class and therefore not the result of final RF classification, an elevated CMP can be seen (Figure 4.8). Contouring Intrusion class membership probabilities at multiple levels from above background level (CMP of 0.125, a random response) to several times background level facilitates a trade-off between correctness and completeness. Many areas exhibit a sharp transition between background levels and an intrusion CMP of 3 times higher than background level (Figure 4.8). Furthermore, the location, form and trend of these zones of elevated CMP better defines the Intrusive lithologies than the final RF classification.

When compared to the FQM geology map, we can see that most intrusions have been captured (Figure 4.8, Figure 4.9), while the false positive rate is low. This result was achieved in areas where no sample of the intrusive class was used in training data, indicating that this method may be viable when using sparse and/or incomplete training data. In this case, intrusions were predicted where no detailed mapping has taken place and their presence in the broader volcano-sedimentary packages is likely. This includes the prediction of discrete but substantial intrusive bodies under regions mapped as glacial till at surface (Figure 4.2, Figure 4.8). Neither mapped nor predicted intrusions show a consistent response to single observables, geochemical or geophysical (Figure 4.9A, 9B), although a small number of units, SIVC for example, can be separated on the basis of magnetic character while a small number of other units are discernible in radiometric data. This highlights the benefit of a machine learning / data fusion approach, making use of many types of data (Figure 4.9C).

The situation where elevated CMPs do not 'win' the final classification is partly due to the overlapping response across class boundaries in training data and extensive adjacent zones. Further contributing factors are the overly smooth response of interpolated data, the sampling interval of geochemical data relative to the width of intrusions and the inherent resolution of the methods themselves, both as a function of the physical and chemical expression of the rocks and the element distribution in eroded and often transported soils. Taking note of CMPs is important as it demonstrates that while the expression of subtle intrusions may not result in a majority decision in the RF classification, they can be detected as a subtler expression in the CMP, i.e. soft classification, metrics produced during the RF classification. As such, we assert the importance of using CMPs in any similar scenario where the goal is the identification of an important rock unit (or any key,

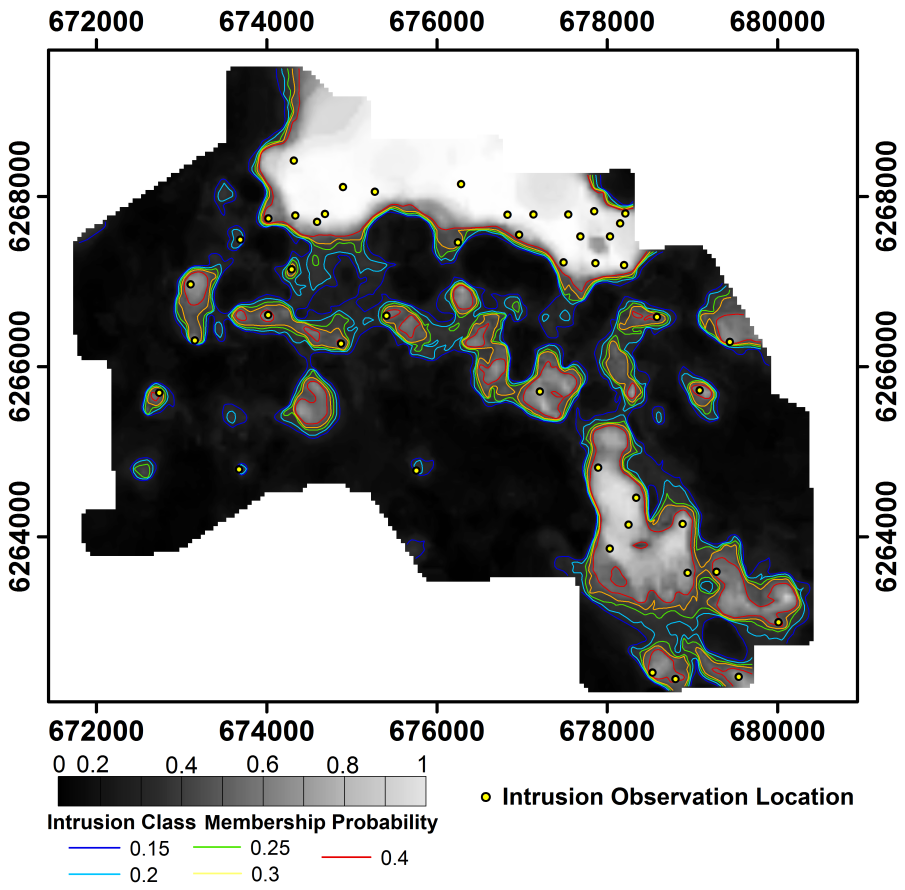


FIGURE 4.8: Intrusion class membership probability. Contours are shown over greyscale image to provide a quantitative level to aid interpretation. The locations of training samples assigned to the Intrusion class, used in classifier training are shown as yellow circles.

indicative feature) as opposed to the overall accuracy of a RF lithology map. These products can be used in conjunction with, or to further guide the deployment of more expensive and labour-intensive mapping, geophysical and geochemical acquisition campaigns.

4.6 Conclusions

Locating intrusions is a key component of exploration for porphyry style Cu Au Mo deposits. In porphyry hosting terrains, such as British Columbia, Canada, where this study is located, lithologies are difficult to discriminate due to similarities in the numerous generations of volcanic, volcanoclastic and intrusive units present, often with similar provenance.

In this study we used systematic rock and soil sampling points as training data points. Additional detailed geological mapping produced a new, FQM geology map, against which the results of our RF classification could be compared. The similarity between many of the lithological classes present resulted in competing class probabilities and an erratic classification. Intrusions, particularly those of a more discrete nature, were inconsistently predicted. This was due to their limited extent, relative to the resolution of the underlying data samples, causing their

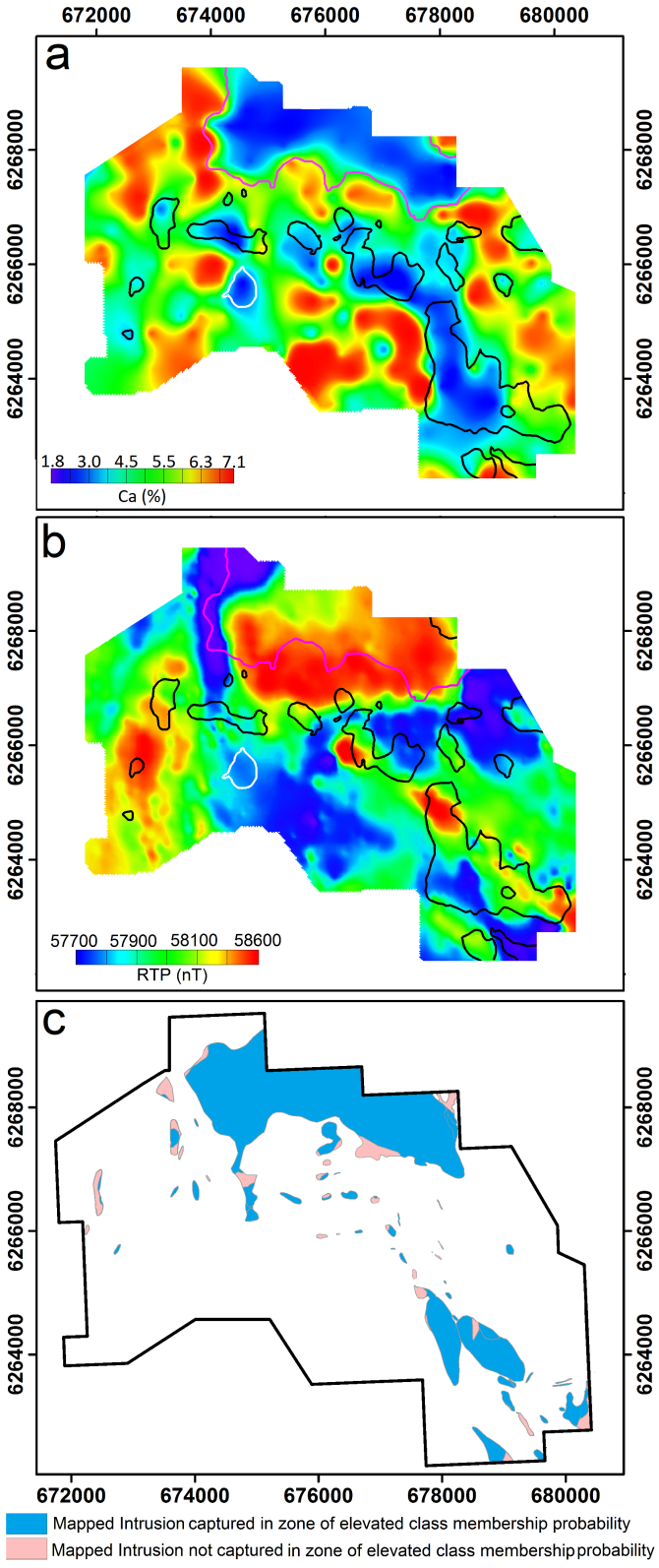


FIGURE 4.9: A and B: Intrusive class membership probability of 0.25 shown over Ca (% Calcium in sample) and RTP respectively. Pink and red polygons are examples of regions of probable intrusions that correspond with Calcium, clearly indicating that classification of intrusions required data other than RTP (and derivatives) C: Region of the project area mapped as intrusions during detailed geological mapping. Mapped intrusions are coloured blue where the Random Forest classification showed an elevated intrusive class membership probability. and pale red where Random Forests did not predict the presence of a mapped intrusion.

expression to be obscured by the spatially larger classes in which they are emplaced. Closer examination of class membership probabilities indicates that there were many locations where the probability of an intrusion being present was significantly elevated above background. Indeed, a large proportion of mapped intrusions were captured by areas of elevated of intrusion class membership probability. Additionally, areas were identified showing an elevated probability of membership to the intrusion class that were not yet mapped in the FQM geology map.

For the task of identification and location of intrusions, Intrusion class membership probability, i.e. not the overall RF lithology map, is the more useful product. Use of this soft classifier has the potential to yield valuable insight, especially at the early stages of exploration. We anticipate that this understanding will find use as a rapid, near real time, tool by exploration teams wishing to predict intrusion locations in order to target and prioritise field activities. More generally, we encourage the use of Random Forest class membership probabilities to gain insight into the occurrence of the classes of greater significance in any data-driven research challenge.

4.7 Acknowledgements

We would like to thank First Quantum Minerals Ltd. for permission to access data. We thank Chris Wijns and Tim Ireland for support and discussion regarding data and results. Stephen Kuhn was supported by a Tasmanian Graduate Research Scholarship (TGRS) from the University of Tasmania. This research was conducted as part of the ARC Industrial Transformation Research Hub for Transforming the Mining Value Chain (project number IH130200004) at the Centre of Excellence in Ore Deposits, University of Tasmania. The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council. We used the Orange software package (Demsar et al., 2013) for RF classification. Pre-processing, interpolation and plotting were performed using Geosoft Oasis Montaj and ESRI ArcGIS.

References

Barnett, C. and Williams, P. (2009). Using geochemistry and neural networks to map geology under glacial cover, *Geoscience bc report, 2009-03,*. accessed 15 November 2017,.
URL: *www.geosciencebc.com/i/project*$_d$*ata/QUESTdata/GBCReport*20093

Bond, G. and Kominz, M. (1984). Construction of tectonic subsidence curves for the early paleozoic miogeocline, southern Canadian Rocky Mountains: Implications for subsidence mechanisms, age of breakup, and crustal thinning, *Geological Society of America Bulletin* **95**: 155–173.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

Breiman, L. (2001). Random forests, *Machine Learning* **45**: 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth Brooks/Cole Statistics/Probability Series, Wadsworth International Group.

Cracknell, M. and Reading, A. (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): 113 – 126.

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

Cracknell, M., Reading, A. and McNeill, A. (2014). Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer – Mt Charter region, Tasmania, using Random Forests^{TM} and Self-Organising Maps, *Australian Journal of Earth Sciences* **61**: 287–304.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. and Zupan, B. (2013). Orange: Data mining toolbox in python, *Journal of Machine Learning Research* **14**: 2349–2353.

Friedman, J. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality, *Data Mining and Knowledge Discovery* **1**(1): 55–77.

Gabrielse, H., Monger, J., Wheeler, J. and Yorath, C. (1991). Part a. morphogeological belts, tectonic assemblages, and terranes, *in* H. Gabrielse and C. Yorath (eds), *Chapter 2 of Geology of the Cordilleran Orogen in Canada, Geological Survey of Canada, Geology of Canada*, Vol. 4, p. 15–28.

Gordee, S., Mortensen, J., Mahoney, J., Hooper, R. and Volcanostratigraphy (2004). Lithogeochemistry and u-pb geochronology of the Upper Hazelton Group, West-Central British Columbia: Implications for Eskay Creek type vms mineralization in Southwest Stikinia, *British Columbia Geological Survey, Geological Fieldwork* p. 311–322.

Grunsky, E. and Kjarsgaard, B. (2016). Recognizing and validating structural processes in geochemical data, p. 85–116.

Harris, J. and Grunsky, E. (2015). Predictive lithological mapping of Canada's north using Random Forest classification applied to geophysical and geochemical data, *Computers and Geosciences* **80**: 9–25.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining*, Inference and Prediction, Springer Series in Statistics, Springer.

Isles, D. and Rankin, L. (2013). Geological interpretation of aeromagnetic data, *Australian Society of Exploration Geophysicists* .

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*, Springer Series in Statistics, Springer.

Johnston, S. (2008). The cordilleran ribbon continent of North America, *Annual Review of Earth and Planetary Sciences* **36**: 496–530.

Kuhn, S., Cracknell, M. and Reading, A. (2016). Lithological mapping via Random Forests: Information entropy as a proxy for inaccuracy, *25th International Geophysical Conference and Exhibition, ASEG, Extended Abstracts*, p. 1–4.

Kuhn, S., Cracknell, M. and Reading, A. (2018). The utility of machine learning in the identification of key geophysical and geochemical datasets: A case study in lithological mapping in the Central African Copper Belt, *First Australasian Exploration Geoscience Conference, Extended Abstracts*, Extended Abstracts.

Kuhn, S., Cracknell, M. and Reading, A. (2019). Lithological mapping in the Central African Copper Belt using random forests and clustering: Strategies for optimised results, **112**: 103015.

Lord, C. (1948). McConnell Creek map-area, Cassiar district, *British Columbia; Geological Survey of Canada, Memoir* **251**: 72.

Lundberg, S. and Lee, S. (2017). Consistent feature attribution for tree ensembles, *Proceedings of the 34th International Conference on Machine Learning*.

Monger, J. (1977). The Triassic Takla Group in McConnell Creek map area, north - central British Columbia, *Geological Survey of Canada* **76-29**: 45.

Monger, J. and Price, R. (2002). The Canadian Cordillera: Geology and tectonic evolution, *CSEG Recorder* **27**: 17–36.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey, *Data Mining and Knowledge Discovery* **2**(4): 345–389.

Nelson, J. and Bellefontaine, K. (1996). The geology and mineral deposits of north-central Quesnellia: Tezzeron Lake to Discovery Creek, central British Columbia, *BC Ministry of Energy, Mines and Petroleum Resources, Bulletin* **99**: 112.

Nelson, J. and Colpron, M. (2007). Tectonics and metallogeny of the British Columbia, Yukon and Alaskan Cordillera, 1.8 ga to the present. Special Publication No. 5, 755-791.

Rebagliati, C., K., B., Bowen, D., Copeland and Niosi, D. (1995). Kemess South and Kemess North porphyry gold-copper deposits, *northern British Columbia, Porphyry deposits of the northwestern Cordillera: Canadian Institute of Mining, Metallurgy, and Petroleum, Special*, Vol. 46, p. 377–396.

Rodriguez-Galiano, V., Chica-Olmo, M. and Chica-Rivas, M. (2014). Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, *Journal of Geographical Information Science* **28**(7): 1336–1354.

Salem, A., Williams, S., Fairhead, D., Smith, R. and Ravat, D. (2008). Interpretation of magnetic data using tilt-angle derivatives, *Geophysics* **73**: L1.

Schiarizza, P. (2004a). Geology and mineral occurrences of Quesnel Terrane, Kliyul Creek to Johanson Lake (94d/8, 9, *Geological Fieldwork 2003, BC Ministry of Energy and Mines, Paper 2004-1*, p. 83–100.

Schiarizza, P. (2004b). Geology of the Kliyul Creek – Johanson Lake area, parts of NTS 94d/8 and 9.

Schiarizza, P. and Tan, S. (2005). Geology and mineral occurrences of the quesnel terrane between the mesilinka river and wrede creek (nts 94d/8, 9), *North-Central British Columbia, British Columbia Geological Survey, Geological Fieldwork* p. 109–130.

Shannon, C. (1948). A mathematical theory of communication, *Bell Systems Technical Journal* **27**: 379–423.

Vucetic, S., Fiez, T. and Obradovic, Z. (1999). *International Joint Conference on Neural Networks of the IEEE. Proceedings*, Vol. 4, p. 2474–2479.

Waske, B., Benediktsson, J., Árnason, K. and Sveinsson, J. (2009). Mapping of hyperspectral aviris data using machine-learning algorithms, *Canadian Journal of Remote Sensing* **35**: 106–116.

Wellmann, J. and Regenauer-Lieb, K. (2012). Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models, *Tectonophysics* **526–529**: 207–216.

Woodsworth, G. (1976). Plutonic rocks of McConnell Creek (94d west half) and Aiken Lake (94c east half) map-areas, *British Columbia; in Report of Activities, Part A, Geological Survey of Canada, Paper 76-1A*, p. 69–73.

Yu, L., Porwal, A., Holden, E. and Dentith, M. (2012). Towards automatic lithological classification from remote sensing data using support vector machines, *Computers and Geosciences* **45**: 229–239.

Chapter 5

Lithological Mapping in the Central African Copper Belt using Random Forests and Clustering: Strategies for Optimised Results

Stephen Kuhn[12], Matthew J. Cracknell[12] and Anya M. Reading[23]

[1] Centre for Ore Deposit and Earth Sciences (CODES) University of Tasmania.
[2] ARC Industrial Transformation Research Hub for Transforming the Mining Value Chain (TMVC), University of Tasmania.
[3] School of Natural Sciences (Maths and Physics), University of Tasmania.

5.1 Abstract

The Trident project is located in the Domes region of the Central African Copper Belt and hosts a number of mineralised systems including the Sentinel (Ni) and Enterprise (Cu) deposits. The project has received extensive systematic geochemical soil sampling in addition to high resolution airborne geophysical coverage. This data-rich environment enables experimentation with machine learning strategies which aim to produce or refine geological maps from limited direct observations.

In this study we present a series of three case studies that test lithological classification using the supervised Random Forests algorithm. These studies inform the situations encountered in mineral exploration including early stage lithology mapping and more mature stage map refinement. We also present a fourth study, using the unsupervised algorithms k-means and Self-Organising Maps, to identify clusters, potentially associated with lithology in absence of *a priori* geological information. Our case studies are most relevant to the situation where the geology of a prospect is largely concealed beneath extensive cover rocks, with some rock types being poorly expressed or even absent in outcrop. We find that sampling from limited outcrop produces a RF lithology prediction that is likely to be incorrect. We demonstrate that balancing sample size through a combination of decimation and bootstrapping can improve results. Additionally, we identify some important indicators in both the predicted geology and uncertainty metrics which could alert an explorer to an inability of their training data to make accurate predictions and to the presence of lithological classes not expressed in outcrop. Sampling from a mature lithology map enables further map refinement and acts as

an objective audit of the existing product. Information entropy (H) is calculated as a metric to describe quantitatively the uncertainty associated with classification, provide valuable information on the geological complexity of the mapped region and highlight areas which are potentially misclassified. Clusters obtained using the k-means algorithm produced a result more consistent with lithology in this instance and was faster; however Self Organising Maps remains attractive due to the production of additional metrics to assess algorithm performance. Clustering could be used either in the development of a first pass interpretation, or in the critical appraisal and subsequent refinement of existing interpretations.

5.2 Introduction

The Trident project, held by First Quantum Minerals Ltd (FQM) is situated in the North-Western province of Zambia (Figure 5.1), a region of the Central African Copper Belt (CACB), one of the world's major mineralised regions, known primarily for copper production, with annual production at 770,598 t as of 2016 (Bank of Zambia, 2016) but also well-endowed with Ni, Cu, Co U, Mo and Au (Selley et al., 2005). The Trident project hosts several major discoveries including the Sentinel (Ni) and Enterprise (Cu) deposits and is located in a region of the CACB which has seen major recent mining and exploration activity with Barrick, Vedanta Resources, Glencore and First Quantum Minerals spending a combined $12.4 billion on new projects between 2000 and 2014 (Mining for Zambia, a Zambia Chamber of Mines Initiative, 2017). The project has received extensive systematic geochemical soil sampling in addition to high resolution airborne geophysical coverage. This data-rich environment enables experimentation with machine learning strategies that aim to predict lithological class and hence produce, or refine, geological maps from limited direct observations.

5.2.1 Geology

The Trident project area is approximately 75 km x 40 km in size and located in the Domes region of the CACB. The region, as described by Capistrant et al. (2015; Figure 1) and references therein, is dominated by the metamorphic basement of the Kabompo dome in the NW and is overlain by the rocks of the Katangan Supergroup, predominately those of the Roan Group, Mwashya Group and Ngumba Group. Large volumes of mafic intrusive units, predominately of gabbroic composition are emplaced in the east. Structurally, the project is dominated by a large NNE trending synform, the hinge of which hosts the Enterprise deposit. The region exhibits a series of NNW striking high-angle faults which crosscut both the basement and overlying Katangan Supergroup. The Domes region is variably subject to greenschist to upper amphibolite grade metamorphism produced during the Lufilian Orogeny (Selley et al., 2005). This heterogeneity is seen locally, within the Trident project area (Capistrant et al. 2015). The stratigraphic positions of some subunits are not well understood due to extensive cover by residual soils with only 0.75% of the area expressed as outcrop. FQM have further subdivided the geology based on in-house mapping and interpretation to produce the initial map of interpreteted lithology used in this study (Figure 5.2).

This map has undergone several updates that combine well-defined stratigraphy, as described above, and lithologies with an unconfirmed stratigraphic position. The

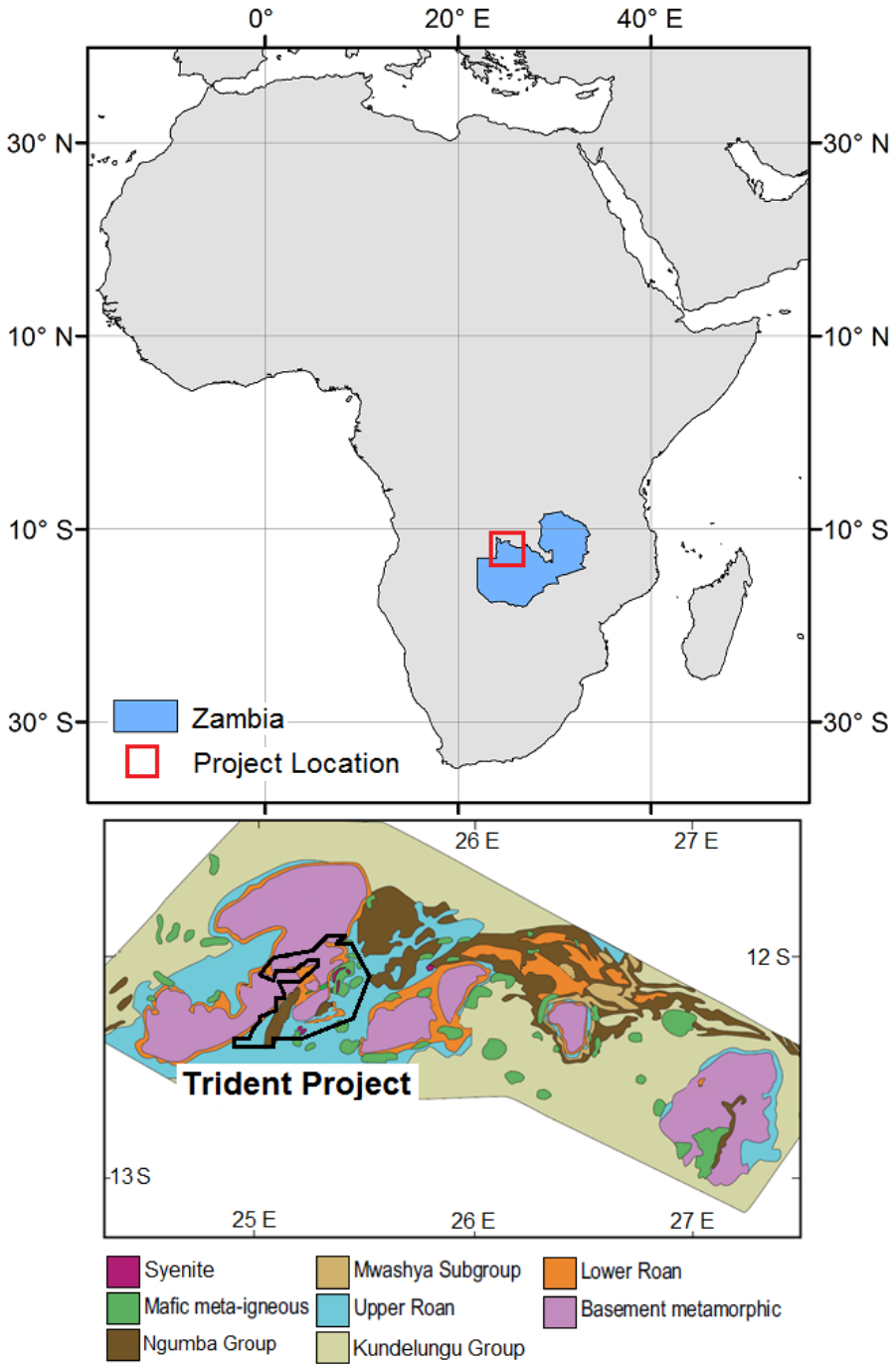


FIGURE 5.1: (Top) Project location relative to the African continent and the country of Zambia. (Bottom) Schematic summary geology of northern North-Western Zambia (modified from Capistrant et al., 2015) showing the location of the Trident project (red outline).

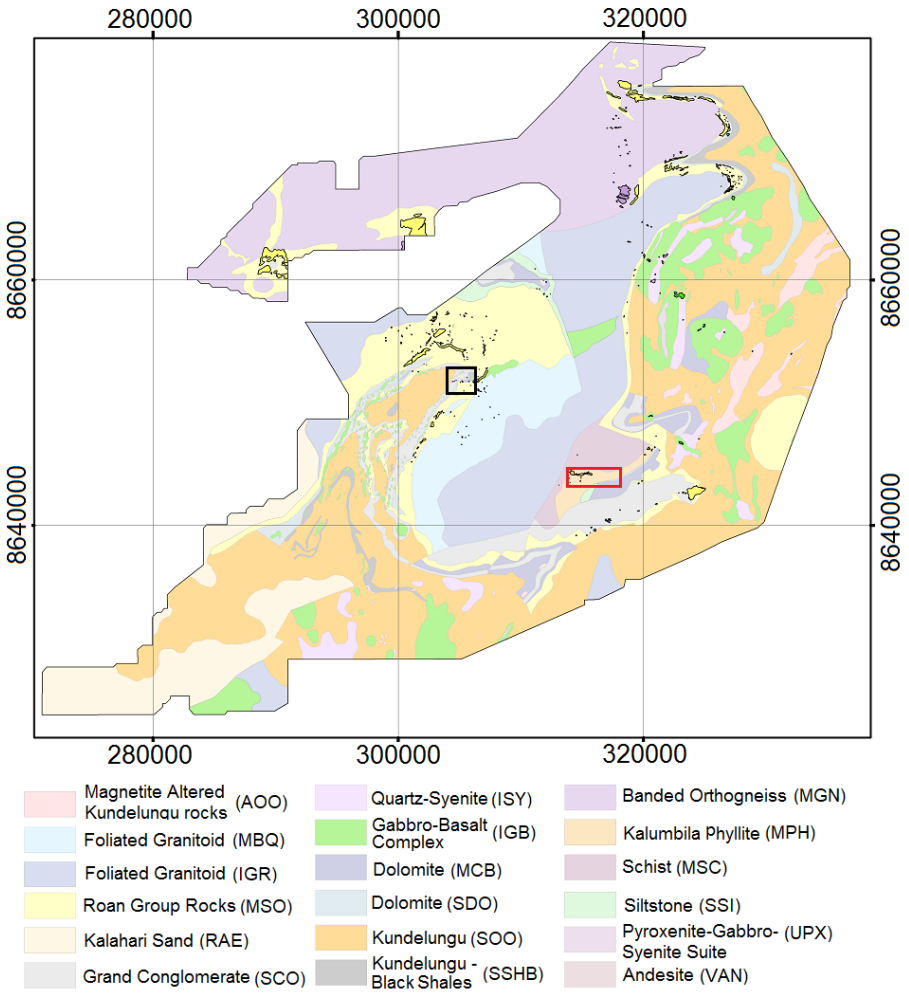


FIGURE 5.2: Initial map of interpreted lithology under cover (pale colours) showing outcrop locations (solid colours). The Enterprise and Sentinel deposits are located within the black and red boxes respectively.

FQM geological interpretation map includes an extensive package of Kundelungu rocks in the east of the project area which is referred to as Upper Roan Group in other work (Capistrant et al., 2015).

5.2.2 Random Forests

Random Forests^{TM} (RF; Breiman, 2001) is a supervised machine learning algorithm (MLA) based on the classification and regression tree method (Breiman et al. 1984). RF, as previously applied described (e.g. Hastie et al., 2009; Kuhn et al., 2016, 2018.) assembles a 'forest' comprising many classification trees (Figure 5.3), each constructed using a unique, random subset of training data. RF compares well to other MLA with regards to accuracy, while remaining straightforward to use and, as such, is considered a good first choice (Cracknell et al., 2014). This is an important consideration for deployment in the geosciences as specialised computing skills may not be available in every exploration team.

RF accuracy is determined by the strength of the classification trees comprising the forest and the correlation between trees (Breiman, 2001). To reduce correlation,

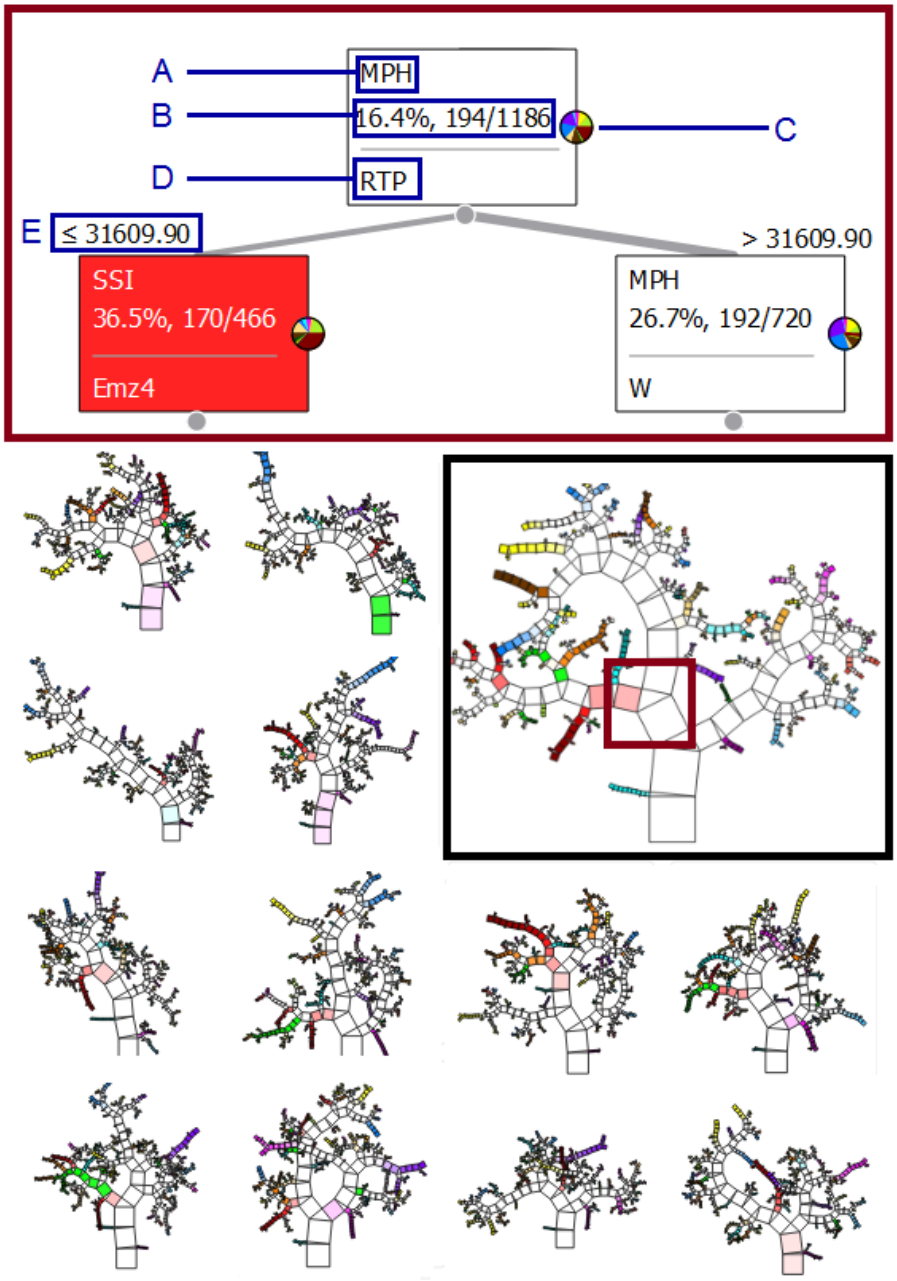


FIGURE 5.3: Schematic example from a RF used in this study highlighting an example of a node split (red box) where A is the nodes dominant class, B is the proportion as percent and count of the node that class occupies, C is the spread of classes also shown as a pie chart, D is the variable used to split the parent node into child nodes, and E is the threshold at which the optimal split in that variable occurred. This node is one of many, from a single unique classification tree (indicated by black box), which is part of a forest (12 examples of 500 shown). Trees are shown as Pythagorean trees (Beck et al., 2014). The relative proportion of parent and child nodes defines the size of squares representing those nodes. Colours note a dominant class, where present.

trees are built on randomly selected subsets of training data (Ta) produced via a process of bootstrap aggregation or bagging (Breiman, 1996). Furthermore, the subset of variables available to split each node in a tree is selected at random. From that subset, the variable which produces the greatest improvement in node homogeneity as defined by decrease in Gini index (Breiman et al., 1984), is selected to split that node (Figure 5.3). Trees are split until homogeneity is achieved or a tolerance is reached. RF classifies each sample by the modal classification of all constituent trees. Accuracy improves with additional trees, until a stable error minimum is reached (e.g. Cracknell et al., 2014; Harris and Grunsky, 2015; Rodriguez-Galiano et al., 2014; Waske et al, 2009). Several studies have applied RF to lithological classification problems. Waske et al. (2009) compared RF with Support Vector Machines (SVM; Vapnik, 1995, 1998) for mapping using hyperspectral imagery. Both algorithms outperformed older classifiers. SVM marginally outperformed RF, however, RF remained an attractive option to the authors due to ease of use.

Cracknell and Reading (2014) compared the performance of RF, SVM, Naïve Bayes, k-Nearest Neighbours and Artificial Neural Networks for geological mapping. They found RF to be most accurate, noting simplicity and lower computational cost as key additional benefits. They found that increasing spatial dispersion of training data improved RF performance, a result which did not manifest to the same extent for other MLAs. Cracknell and Reading (2014) also compared RF and SVM for mapping and identification of geological boundaries; and zones of structural complexity. They concluded that both RF and SVM were similarly accurate while RF produced more meaningful results with high RF uncertainty associated with map boundaries and complex regions. These findings were reproduced by Kuhn et al. (2016), who also noted a relationship between uncertainty and map inaccuracy. Cracknell and Reading (2014) successfully used RF to refine geological mapping in western Tasmania, subsampling a geological map as training data. Harris and Grunsky (2015) used a similar approach in northern Canada, using lake sediment samples and field observations to train RF, again noting the value of RF as a first-pass mapping tool. Kuhn et al. (2018) deployed RF in a reconnaissance setting in the Eastern Goldfields of Western Australia, refining a geological map using geophysical data and highlighting the applicability of uncertainty in assessing map validity.

5.2.3 Quantification of uncertainty

RF classifies each sample by majority vote cast by all component decision trees, however, a more detailed distribution of probabilities exists for each possible class. Class membership probabilities are recorded, defining the proportion of trees that voted for each class (Hastie et al., 2009). Individual class probabilities can be assessed in isolation or the probability distribution can be quantified as a single number. In this study, as a proxy for uncertainty, we use information entropy (H; Shannon, 1948) defined as:

$$H = -k \sum_{i=1}^{n} p_i log p_i$$

were p_i is the class membership probability at location i, n is the number of candidate classes, k is a positive constant. Both k and the logarithm base are arbitrary and are used to manage scale. H describes the level of disorder in a

system. A minimal value corresponds to complete homogeneity and a maximal value corresponds to equal possibility of all classes. H preserves monotonicity. Increasing the number of candidate classes produces a higher possible H.. H has proven effective in defining the spatial distribution of uncertainty (Wellmann and Regenauer-Lieb, 2012; Kuhn et al., 2016). Values can be normalised (H_{norm}) for the number of candidate classes. H_{norm} represents the minimum to maximum possible H for each sample, allowing samples to be compared with regard to how closely each approaches its own maximum possible H. For example, a sample with two possible and equally probable classes; and another with five possible and equally probable classes; will each produce H equal to one. H responds to complexity: the number of classes possibly interacting at a given location. H_{norm} is more closely associated with predication inaccuracy (Kuhn et al., 2016). It is important to note the distinction between inaccurate mapping and predictions that are inconsistent with the starting interpretation map does not discount the possibility the interpretation was incorrect, and RF has identified the correct classification. The behaviour of H and H_{norm} may provide insight into whether this has occurred (Kuhn et al., 2016; 2018).

5.2.4 Clustering

The k-means algorithm (Lloyd, 1957; 1982) is a widely used clustering algorithm that operates on the principle of partitioning data based on similarity (Macqueen, 1967) The k-means algorithm is a pragmatic first choice for geoscientific applications due to conceptual and operational simplicity. The k-means algorithm starts with the random placement of a given number of centroids in the data space. Euclidean distance to each data point is calculated and each data point assigned to the nearest mean, dividing the dataspace via Voronoi partitioning. Subsequent iterations calculate new means using all data assigned to each centroid and centroids are adjusted to those positions. This process is repeated until centroid adjustment does not result in further re-assignment or until an iteration cap is reached. As implemented in this study, silhouette analysis (Rousseeuw, 1987) provides a measure of dissimilarity for points within clusters, as compared with dissimilarity to the nearest neighbouring cluster. This facilitates an objective selection of number of clusters needed to produce best separation between clusters. Random seeding of starting centroids can produce high processing times and convergence to local error minima. The k-means++ algorithm (Arthur and Vassilvitskii, 2007) controls seeding of starting centroids and produces superior processing performance and accuracy than random seeding. All further reference to k-means in this paper relate to k-means with k-means++ seeding.

Self-Organising Maps (SOM), developed by Kohonen (1982; 2001), maps high dimensional data onto a lower dimensional plane in such a way that preserves the topological relationships in the dataset (Penn, 2005). A map is defined, with a number of nodes relative to the number of input data. Data are treated as n-dimensional vectors. Vector similarity between data and nodes are measured and winning nodes updated to better resemble the assigned data, as are those within a defined radius of a winning node, by a percentage of that applied to the winning node. The process is repeated, with the radius of influence and percentage of modification reduced iteratively. SOM has been deployed in the geosciences (e.g. Fraser and Dickson, 2008; Berlein et al., 2008; Cracknell, Reading and McNeill, 2014; Cracknell et al., 2015) with useful clustering results and visual outputs such as the

unified distance matrix (Ultsch and Vetter, 1994). In this study, complete linkage hierarchical clustering (Defays, 1977) is used for additional cluster reduction with optimal cluster number assessed using the Davies-Bouldin index (DBI; Davies and Bouldin, 1979). The method of complete linkage reduction of SOM clusters will be referred as SOM-CL in this study.

5.2.5 Objectives

We conduct four experiments that simulate geological mapping using machine learning for a variety of input conditions. Two of these studies describe the use of RF for mapping using samples from outcrop, both on an "as is" basis (replicating an early stage in exploration) and balanced for class sample size. A third study uses RF to reclassify the project using a small subset of training data, sampled at random from a company interpretation map. The goal of the third study is to assess the viability of RF to audit objectively and, where possible, improve upon an existing map (replicating a more mature stage in exploration). Lastly, we assess the ability of the clustering algorithms to produce a classification, in the absence of any user input, which corresponds to mapped geology at the scale of the project.

5.3 Data and Methods

5.3.1 Data compilation and pre-processing

Data used in this study were provided by FQM. These comprise both geophysical and soil geochemical data (Table 5.1; Figure 5.4). Additional geophysical datasets were derived from those provided and the Shuttle Radar Topography Mission (SRTM; National Aeronautics and Space Administration, 2006) digital terrain model (DTM) were added.

TABLE 5.1: Variables remaining after the removal of highly correlated variables.

Geophysics		Soil geochemistry		
Dataset	Abbreviation	Ag	Fe	Se
Reduced to Pole Total Magnetic Intensity	RTP	Al	In	Sn
RTP - First vertical Derivative	RTP_1vd	As	La	Sr
Total Magnetic Intensity – Analytic Signal	ASIG	Au	Li	Ta
Radiometric – Potassium	K_rad	Ba	Mg	Te
Radiometric – Thorium	Th_Rad	Be	Mo	Ti
Radiometric – Uranium	U_Rad	Ca	Na	Tl
		Cd	Ni	W
Airborne Electromagnetic Channel 4	Emz4	Co	P	Y
(150 ms): z component		Cr	Pb	Zn
		Cs	Re	Zr
DTM (Shuttle Radar Topography mission)	DTM	Cu	S	

Soils in the project area are believed to be residual, and hence, reliable proxies for the lithologies below. Geochemical data with values of 0 or below detection limit were assigned by default, a value equal to half the detection limit of that element. Aeromagnetic (flown at 100 m line spacing) and airborne electromagnetic data (flown at 200 m line spacing) were gridded using minimum curvature at one fifth and one quarter of their respective flight line spacing (20 m and 50 m cell size

respectively). Geochemical data (sampled at 300 x 300 m) were gridded to a 100 m cell size. All data were resampled to a regular grid of 100 m x 100 m and compiled into a matrix taking the form of: x, y, p_1, p_2, ..., p_n , where x and y are coordinates and p are values of each variable at a given sample location. This database comprises approximately 178,000 instances, each with 59 variables, and was used to partition training and test subsets for the RF experiments.

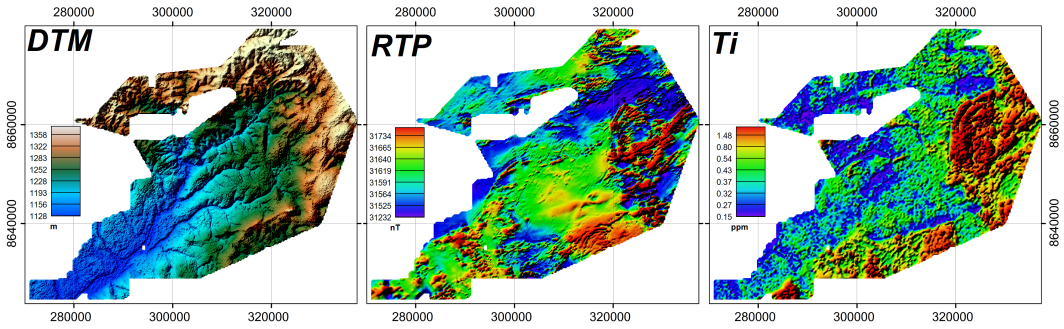


FIGURE 5.4: Examples of 3 variables used in this study: DTM, RTP magnetics and Ti. These variables were deemed useful in case studies C1, C2 and C3.

5.3.2 Removal of highly correlated variables

High correlation between variables suggests that they are not independent and are duplicating information. This can lead to supervised classifiers placing undue emphasis on those features (Guyon, 2008). Where a pair of variables exhibited a high correlation, defined as those with a Pearson's correlation coefficient >0.8, one of those variables was removed. In cases where a variable exhibited excessive noise, or a large number of below detection limit or missing samples, that variable was removed. A total of 15 variables were removed, reducing the number of variables for consideration to 44 (Table 5.1).

5.3.3 Variable ranking

Previous studies (e.g. Cracknell et al., 2014; Kuhn et al., 2016) have shown that a point of diminishing returns exists, beyond which additional variables do not improve accuracy and unduly complicate the interpretation of results. RF has an inherent mechanism for ranking variables, (Breiman, 2001). Each variable is permuted and the change in accuracy measured. Variables are ranked from highest to lowest importance, with those that the classification accuracy is most sensitive, deemed most important. Variables were successively added in rank order in addition to those prior (i.e. 1, 1+2, 1+2+3 and so on), and accuracy tested by 10-fold cross-validation. Variables were added until no further improvement was reached. This was defined as the last instance where the addition of a variable produced a change in cross validation accuracy of 1%. Variable ranking is specific to the training data used. Rankings were produced in this manner, independently, for each of three RF case studies.

5.3.4 Sampling

Case study 1

Case study one (C1, early exploration stage) used mapped outcrop locations as training data (Figure 5.5A). Samples were treated on an "as-is" basis with sample size controlled by the abundance of each lithology in outcrop. This resulted in highly imbalanced training set sizes, favouring the Roan Group and Banded Orthogniess rocks (Figure 5.5A). This is not an optimal training set as RF produces the best results when class sample sizes are balanced, otherwise it is prone to over-fitting to classes with more samples. Outcrop observations do not represent all lithologies (12 of 17 represented) and is restricted to the east of the project and yields different training sample sizes. Our objective in using this raw sample set is to investigate resulting errors in map outputs and uncertainty. We also investigate how such errors might be identified in the absence of a priori knowledge of the extent to which outcrop reflects geology undercover and/or without known geology with which to verify results.

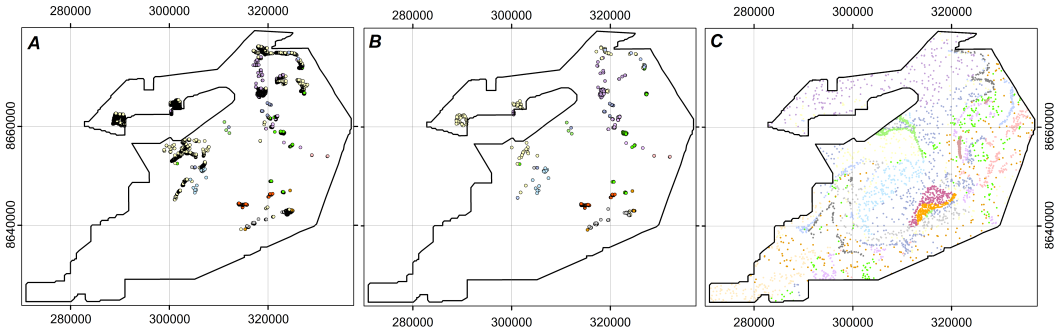


FIGURE 5.5: Training data locations for (A) case study C1, (B) C2 and (C) C3. Note the diameter of each sample in (A) and (B) has been increased by a factor of 7 and in (C) by a factor of 3, for legibility. See Figure 2. for lithology colour key.

Case study 2

Case study two (C2, early stage with method refinement) started with the C1 training set. In order to rectify the imbalance in sample size in C1, we used a combination of bootstrap sampling (Hastie et al., 2009) and decimation. Sample sizes of 50, 100, 200 and 400 were investigated to find the balance between preserving real samples and introducing artificial samples (Table 5.2). A sample size of 100 per class was deemed to provide this balance of adequate sample size while introducing an acceptable number of synthetic samples (Figure 5.5B). Larger sample sizes retained more real data but introduced an unacceptably high proportion of synthetic samples across all represented classes.

Case study 3

Case study three (C3, mature exploration stage) investigates the deployment of RF at more advanced exploration project maturity than C1 and C2. As such this study capitalises on much more extensive geological information in the form of a well-developed company geological interpretation map. The objective, rather than using outcrop to predict geology in unmapped regions as in C1 and C2; is to refine

TABLE 5.2: The decimation and resampling used for balanced training classes of various sizes. A smaller class requires the least introduction of bootstrapped samples however a large number of real data are excluded. A larger class makes better utility of real data however the numbers of bootstrapped data are excessive. 100 samples per class represents an optimal balance between use of real data and introduction of bootstrapped samples.

Class Size	Decimate (D)	Bootstrap (B)	D to B Ratio
50	72	19	0.26
100	56	49	0.88
200	35	121	3.46
400	7	278	39.7

the existing geological interpretation. Additionally, through the calculation of H, we will provide insight into map regions defined by geological complexity while providing an indication of areas with a high probability of incorrect classification. A stratified, spatially balanced random sample was taken from the FQM geological interpretation map (Figure 5.5C). In this case, 200 samples per class were taken from each of the 17 mapped lithological classes. The remainder of the dataset was held for testing, unseen by the classifier.

Case study 4

Case study four (C4, clustering approach) tests the use of the k-means and SOM algorithms to define natural groups in the data, i.e. without the introduction of user input or influence resulting from the use of training data or predefined classes. This has the advantage of being able to identify features not represented in the training data. The disadvantage however is that there is no control over the correspondence of clusters to lithology or other geological phenomena such as alteration. Nevertheless, at scale of this project, the geology comprises several distinct domains. This study seeks to test whether clustering is a viable means of producing a first-pass interpretation map in a situation akin to C1 and C2. To address the relative magnitude of datasets, all variables were normalised such that the mean has a value of 0 and a variable at one standard deviation from the mean has a value of 1. The complete database of approximately 178,000 samples was used. A number of iterations were tested for each clustering exercise. For k-means with 20 clusters, the upper bound for the number of clusters allowed in this study, 99.1% of samples were partitioned into their final clusters after 300 iterations (Figure 5.6). As such 300 iterations were used for all k-means models. SOM parameters including map size and dimensions, were investigated and a 45 x 45 node map used in this study. Both algorithms were tested using all variables below the 0.8 correlation threshold and again using those variables ranked most important during C2, representing the optimal understanding of variables from outcrop mapping alone.

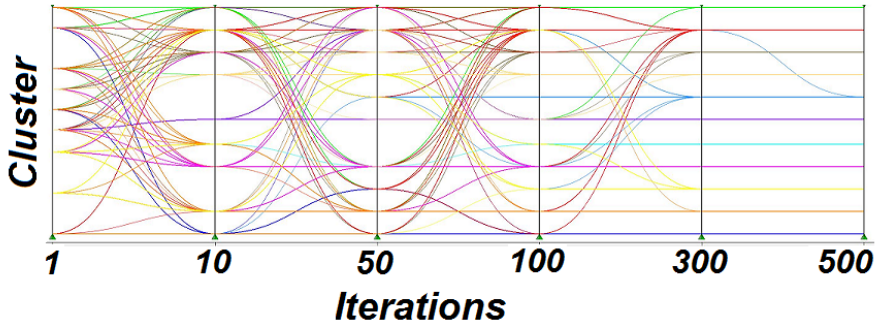


FIGURE 5.6: k-means convergence vs iterations performed. Lines represent the assignment and subsequent reassignment/refinement of samples as the number of iterations is increased. Lines are smoothed between experiments and reflect the reassignment path (and not the assignment of samples at iteration increments between those displayed).

5.4 Results

5.4.1 Ranking and variable selection

A 500 tree RF was used to rank variables in the C1, C2 and C3 training datasets. The C1 training data produced a peak cross validation accuracy of 75.4% using the top 9 ranked variables (Table 5.3; Figure 5.7). Ranking of training data from C2 defined 10 relevant variables (Table 5.3; Figure 5.7), producing a peak cross validation accuracy of 88.8%. Ranking of datasets using training data from C3 identified 12 relevant variables (Table 5.3; Figure 5.7), producing a peak cross validation accuracy of 80.5%.

TABLE 5.3: Ranking, variable (Var), RF score (RF) and 10 fold cross validation accuracy (Acc) for C1, C2 and C3, shown to a depth of 15 variables. Note that the cross validation accuracy refers to the result obtained with the use of a given variable in addition to those ranked higher. Green indicates the optimal cut off for variables used in each case.

	C1			C2			C3		
Rank	Var	RF	Acc %	Var	RF	Acc %	Var	RF	Acc %
1	DTM	6.8	32.3	DTM	5.2	70.4	DTM	7.3	23.4
2	Cd	4.8	50.1	As	4.0	77.1	EMZ4	7.0	42.4
3	Ag	3.8	60.6	Th_rad	3.4	78.2	Ti	3.9	58.4
4	RTP	3.7	67.4	Ta	3.3	81.5	RTP	3.6	67.2
5	Se	3.0	70	Mg	3.1	83.8	Cu	2.9	70.4
6	Ars	2.9	71.5	Ti	3.1	84.8	Cd	2.8	73.1
7	Pb	2.8	72.9	Emz4	3.0	86.8	ASIG	2.7	75.5
8	Emz4	2.7	73.9	Ni	2.8	87.1	In	2.7	77
9	Ti	2.7	75.4	La	2.8	86.9	Mg	2.6	78.1
10	Fe	2.6	74.5	RTP	2.6	88.8	Fe	2.6	79
11	In	2.5	75	Te	2.6	88.1	Ba	2.6	79.5
12	Sr	2.4	75.3	Mo	2.5	88.8	W	2.6	80.5
13	Ni	2.3	75.6	Cd	2.5	89.5	Zn	2.6	79.8
14	Cr	2.3	75.8	Cr	2.5	89.4	Ni	2.6	80.1
15	Cu	2.2	75.9	Ca	2.5	89.3	Ta	2.6	79.9

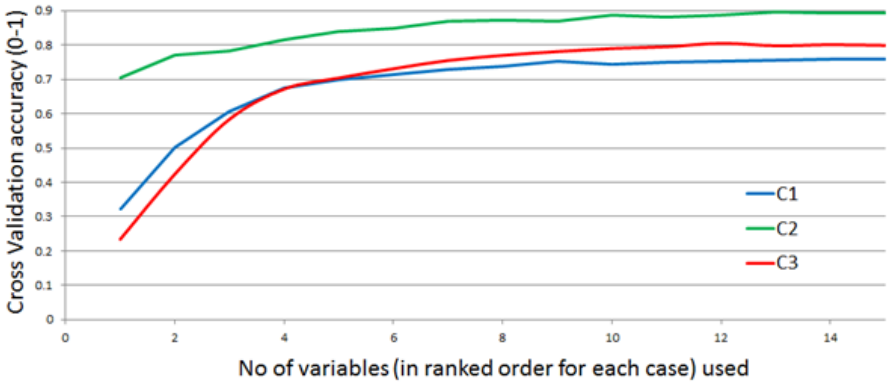


FIGURE 5.7: Cross validation accuracy with addition of successively lower ranked variables for each RF case study. (C1) sampling from outcrop, (C2) class size balanced sampling from outcrop and (C3) sampling from a geological map. Note the accuracy using balanced outcrop-based sampling (C2) is strongly influenced by overfitting of the RF model to a small and more homogeneous dataset which does not well describe the full variability of those units were the whole unit available for sampling.

5.4.2 C1 Classification results

Prediction of lithology using outcrop led to a training sample imbalance in favour of the MSO and MGN classes. This resulted in a RF model dominated by the MSO and MGN classes (Figure 5.8A). A pixel by pixel comparison showed results of this case study to be consistent with the interpreted geology map in only 17% of cases. H indicates high uncertainty within the area represented by outcrop with a region of low H in the south west (Figure 5.8B).

Extending beyond the training data to the southwest, is a zone of low H, as observed in C1. Mapping shows a correlation with terrain and drainage patterns (Figure 5.9A; DTM in Figure 5.4). We assert that the high rank and thus influence of the DTM, while in part due geological controls on topography is also due to the positions of outcropping samples serving as a proxy for geographic location. This may not conform with the range of elevations occupied by that class across non-outcropping areas and thus biases the classification in favour of the particular elevation at which training data were observed. The omission of the DTM resulted in a lithology prediction that saw better recovery of interpreted boundary geometries (Figure 5.9B) in the areas well represented by training data; and better prediction of gabbros in the south of the map. H in this case was higher across the project (Figure 5.9D) than was the case for classification results produced with the DTM included and showed a more chaotic spatial distribution and relationship with lithological boundaries.

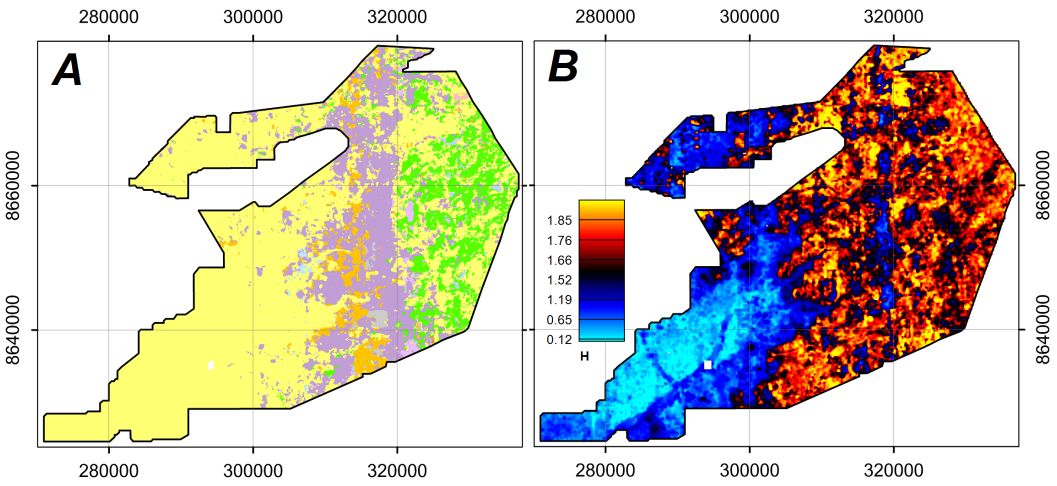


FIGURE 5.8: (A) Classification output using C1 training data. See Figure 2 for lithology colour key. (B) H associated with C1 classification output. Note that in addition to poor accuracy with respect to interpreted lithology on a pixel by pixel basis, interpreted geometry and structure are absent, in favour of broad N-S trending domains. Anomalously low H associated with extrapolation of nearest sampled lithology into the south west is a warning that training data do not represent lithologies in that region and assumptions regarding the behaviour of uncertainty (Cracknell and Reading, 2014, Kuhn et al., 2016) are not valid.

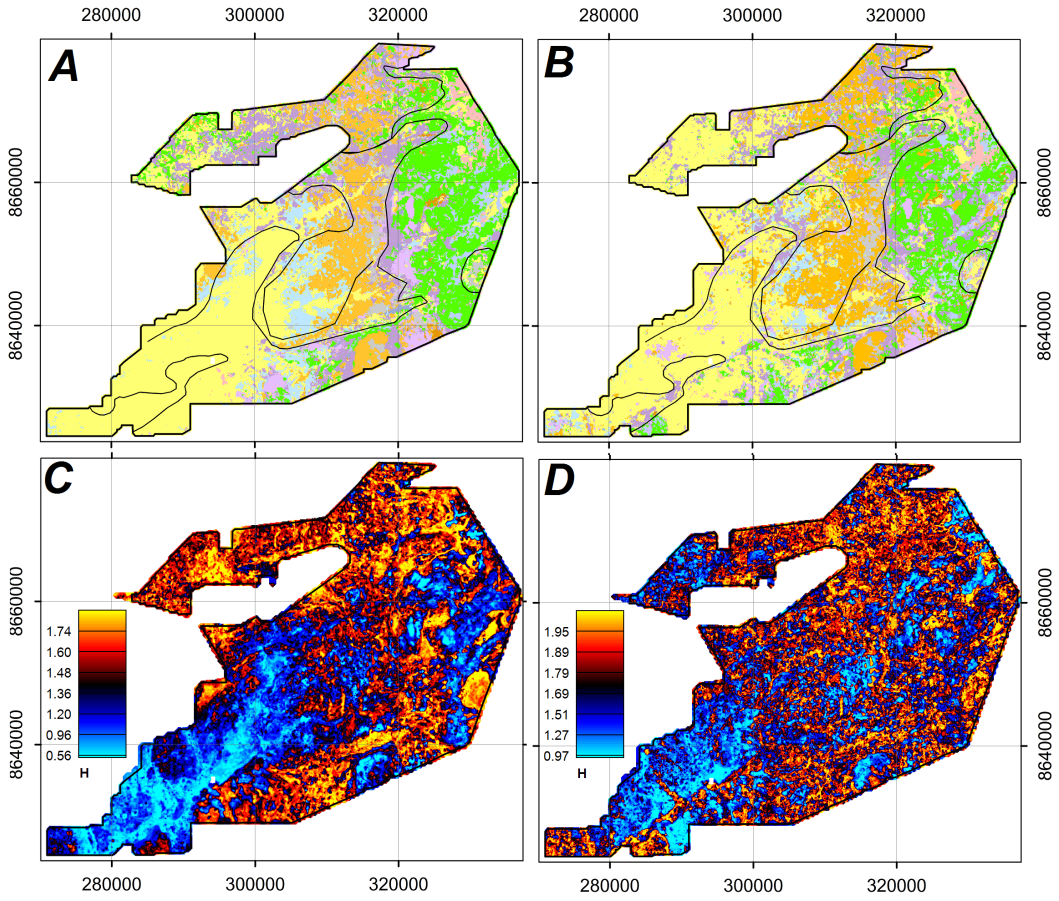


FIGURE 5.9: (A) Classified lithology map refined using C2 training data. See Figure 2 for lithology colour key. (B) Classified lithology map using C2 training data adjusted to omit the DTM. (C) H associated with (A). (D) H associated with (B). Note that while lithology prediction accuracy is poor on a per pixel basis, major geometries/boundaries are present.

5.4.3 C3 Classification results

This case study (C3) made use of a well-developed geological interpretation map for the generation of training data. RF produced predictions (Figure 5.10A) with 67.2% consistency with respect to that map (Figure 5.10B). The confusion matrix associated with this classification (Table 5.4) shows that 10 of the 17 lithological classes achieved a recall in excess of 75% and a further three classes above 65%.

TABLE 5.4: Confusion matrix. Red, Orange and Blue text represent < 60, > 60 and > 75 percent of samples classified consistent with the interpreted geology map. Prediction consistency is expressed as a percentage and the relative size of classes given as number of samples. Rock codes are as per Figure 5.2.

		Predicted																
Map	AOO	IGB	IGR	ISY	MBQ	MCB	MGN	MPH	MSC	MSO	RAE	SCO	SDO	SOO	SSHB	SSI	UPX	
AOO	**88**	4	0	2	0	3	0	0	0	0	0	0	0	2	0	0	0	9121
IGB	10	**62**	2	6	1	3	0	0	0	0	1	3	2	4	2	0	4	12514
IGR	0	1	**68**	0	13	1	0	0	7	3	0	2	1	3	0	0	0	17459
ISY	3	11	0	**79**	0	3	0	0	0	0	0	0	0	3	1	0	0	8300
MBQ	0	0	7	0	**81**	0	0	0	5	2	0	3	0	0	0	1	0	11562
MCB	2	2	0	2	0	**79**	0	1	2	1	0	5	0	2	0	3	0	10355
MGN	0	0	1	0	0	0	**96**	0	0	3	0	0	0	0	0	0	0	28095
MPH	0	0	0	0	0	2	0	**91**	3	0	0	1	0	0	0	3	0	1352
MSC	0	0	1	0	2	0	0	11	**85**	0	0	1	0	0	0	0	0	4513
MSO	1	1	5	0	6	2	10	0	2	**54**	1	3	5	2	3	5	0	14030
RAE	0	1	0	0	0	0	0	0	0	0	**95**	0	2	1	0	0	0	14795
SCO	0	1	2	1	5	13	0	0	3	7	3	**52**	4	4	2	2	0	7076
SDO	0	2	0	0	2	0	0	0	0	7	3	3	**67**	2	13	0	0	5699
SOO	10	7	1	9	2	11	1	0	0	1	7	3	4	**38**	5	0	0	21049
SSHB	1	3	0	1	0	0	0	0	0	1	12	0	5	2	**74**	0	0	5379
SSI	0	0	1	0	0	2	0	1	1	2	0	0	0	0	0	**92**	0	2589
UPX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	594
Count																		

Bulk inaccuracy was largely a function of the undifferentiated Kundelungu rocks being partitioned into other, adjacent lithologies, many of which were more concisely defined subunits of the Kundelungu Group (Table 5.4). This can be seen clearly in the expression of class membership probabilities, examples of which are shown in Figure 5.11. The MGN lithology class (Figure 5.11E, predicted with a recall of 96%, exhibits very concise and distinct regions where this class was probable.

Conversely, the lithology class SOO (Figure 5.11D) was not predicted by a large majority, as shown by relatively low class membership probabilities across its spatial range. H (Figure 5.12A) highlights areas of geological complexity and shows a relationship with lithological contacts. This is prevalent in the centre and south of the project. H_{norm} (Figure 5.12B) shows a larger proportion of cells internally approaching maximal possible H and demonstrates a correlation with those cells which were classified inconsistently with the starting geological interpretation map (Figure 5.13).

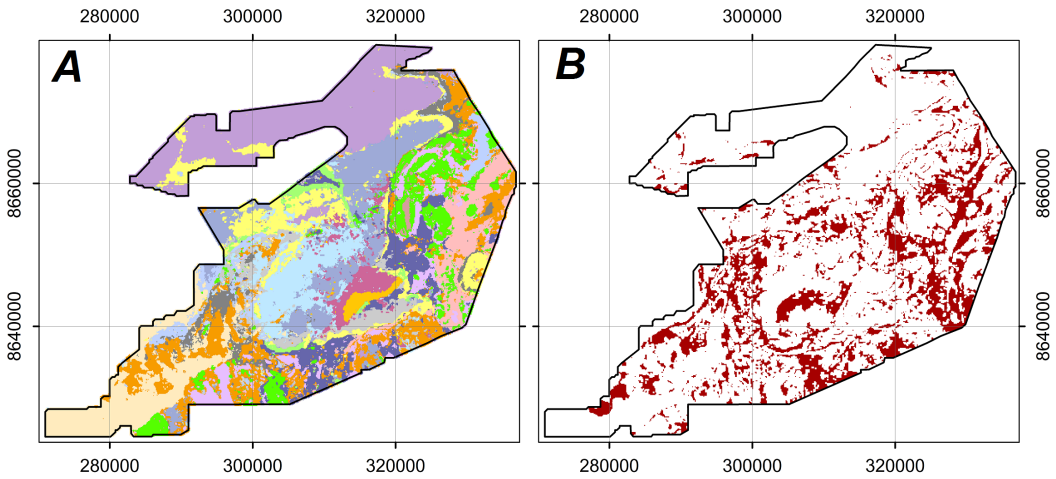


FIGURE 5.10: (A) Classified lithology map refined using training data C3. See Figure 2. for lithology colour key. (B) Comparison with the initial map of interpreted geology (Figure 2) as consistent (white) and inconsistent (red).

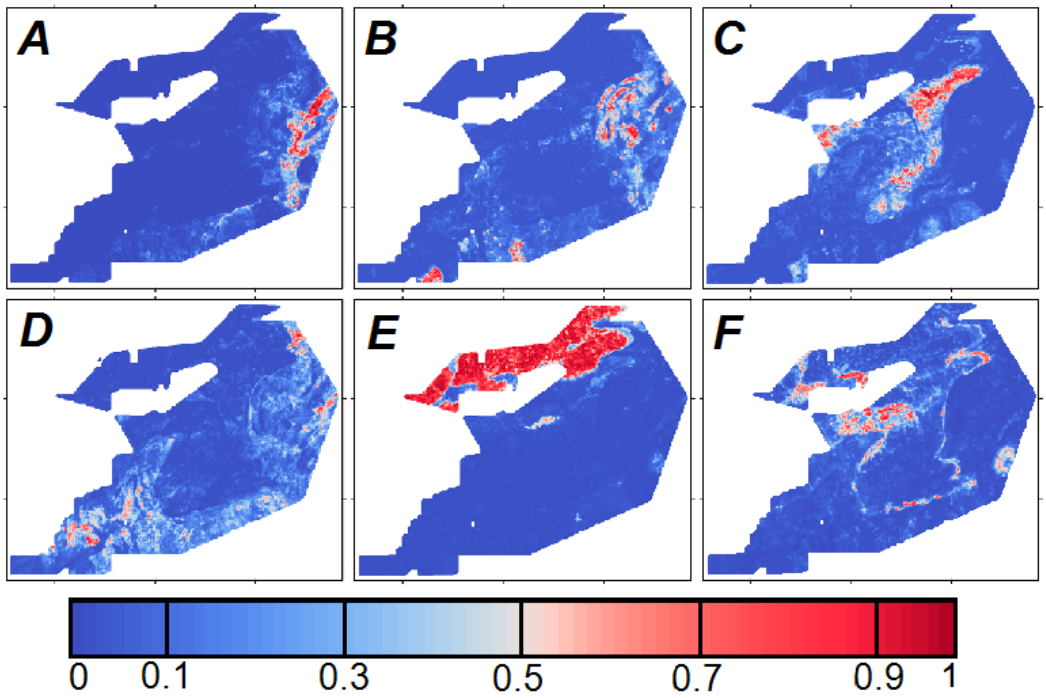


FIGURE 5.11: Examples of case C3 class membership probabilities. (A) AOO, (B) IGB, (C) IGR, (D) SOO, (E) MGN and (F) MSO. Rock codes are given in Figure 5.2.

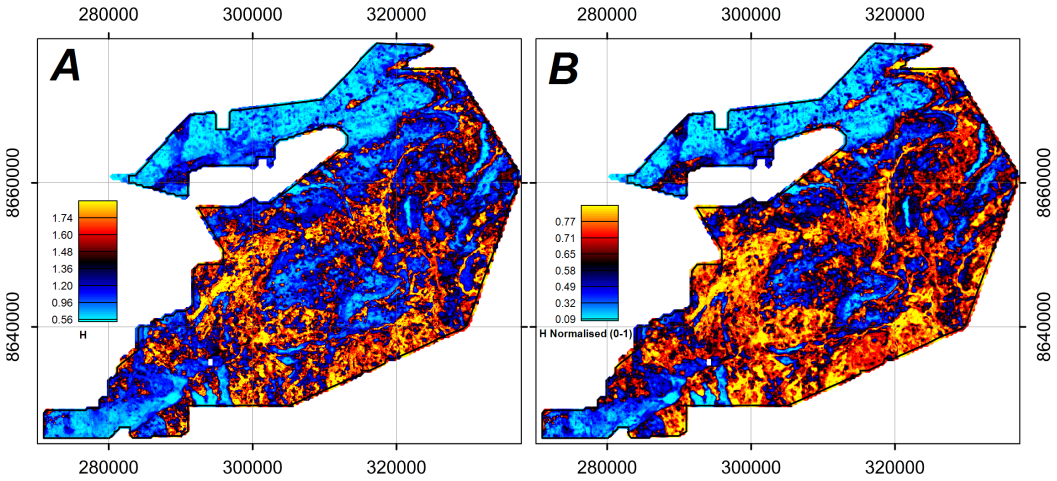


FIGURE 5.12: Case study C3: (A) H and (B) H normalised to 0-1. H is an indication of the disorder at a given point and rises as complexity, i.e. the number of possible classes, increases. When normalised, H provides an indication of how closely a given pixel reaches its maximum possible state of disorder. As such, pixels can be compared and can be a better proxy for prediction accuracy.

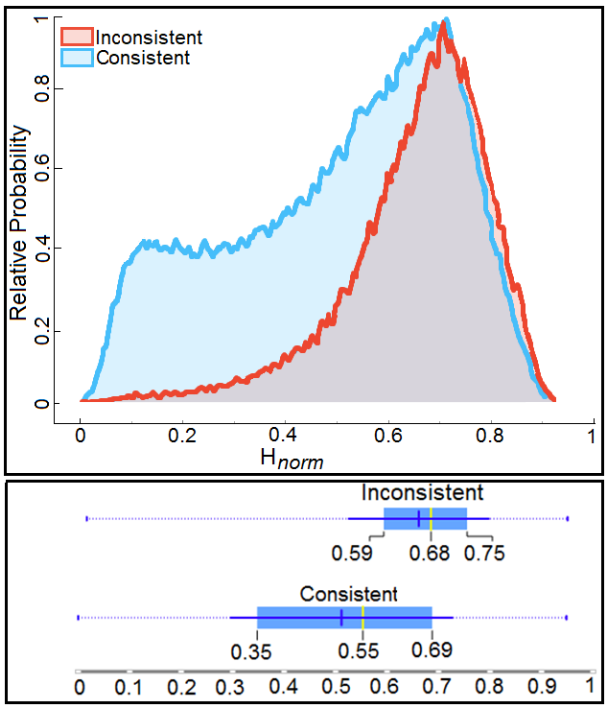


FIGURE 5.13: The distribution of H for C3 partitioned into two groups: samples classified consistently, or inconsistently, relative to the initial interpreted lithology map (Figure 2). (Top) The relative probability of a consistent or inconsistent classification for any given Hnorm. (Bottom) Box plot showing the distribution of Hnorm for consistent and inconsistently classified sample populations. Note that at above a Hnorm of 0.75, there is a greater probability of encountering an inconsistent classification than consistent however there is considerable overlap from 0.6 to 0.75 where either is similarly probable. Below a Hnorm of 0.5, a consistently classified sample is considerably more probable.

5.4.4 C3 Clustering results

Using k-means and SOM-CL produced a series of outputs of 2 to 20 clusters with an optimal cluster number defined by silhouette and DBI (for k-means and SOM-CL respectively). When using all datasets, both methods showed a strong relationship with drainage patterns, regardless of cluster numbers. As such, both methods were performed using only those elements ranked as non-redundant by RF in C2. Outputs for k-means and SOM-CL optimised at 3 and 5 clusters respectively. In both cases, this reflected a separation of the Kundelungu Group from the metamorphic basement. Based on the number of lithological classes expressed in outcropping geology, information that would be available at the earliest stages of a project, we further constrained cluster number to between 10 and 20. With this limit in place both k-means and SOM-CL defined the optimal number of clusters as 11. Both k-means and SOM-CL results showed a strong spatial resemblance to the interpreted geology map of the project (Figure 5.14).

5.5 Discussion

5.5.1 Ranking of input data

Dataset ranking is a necessary component of RF classification while also providing a rapid and objective means of prioritising data for other areas of geological and geochemical investigation. The sample used for classification in C1 produced spurious results due to the large imbalance in class size. Conversely, ranking using a properly balanced sample (C2) produced a set of relevant datasets which would prove insightful geochemical interpretation (discussed in detail in Kuhn, et al., 2018). This set (Table 5.3) included widely used geophysical mapping datasets (RTP, EMZ4) and several high field strength elements that are well known lithological discriminators such as Ti and Ta (Pearce and Norry, 1979; MacLean and Barret, 1999). That these datasets, well known for use in conventional geological mapping, were prominent in RF ranking lends credence that the RF assessment of ranked datasets was geophysical and geochemically sound; providing confidence in the RF classification and other interpretations based on these findings. In addition to these well-known datasets, others were included, the importance of which may be idiosyncratic to the project. La, for example, trends through the central-northeast of the project. Company geologists (Ireland, pers. comm., 2016) identify this feature as a monazite trend. Other elements such as As and Mg, also ranked as necessary by RF, have been used by company geochemists for the subdivision of mafic packages and partitioning of talc rich rock units respectively, further demonstrating that RF rankings are geologically meaningful.
Ranking of datasets using Ta, sampled from a geological interpretation map (C3) saw the increased prioritisation of geophysical datasets, with the EM and RTP datasets featuring at second and fourth most important, respectively (Table 5.3). This is consistent with the additional information used in producing this map, as compared to a model comprising purely observations. The prominence of these datasets is likely a reflection of their use in defining lithological zones during geological interpretation.

The RTP magnetics dataset was ranked as necessary in all cases, while the first vertical derivative (1VD) was redundant. We assert that at the scale of mapped lithology, the 1VD, a high-pass filter, is responding to sub-units or other textures

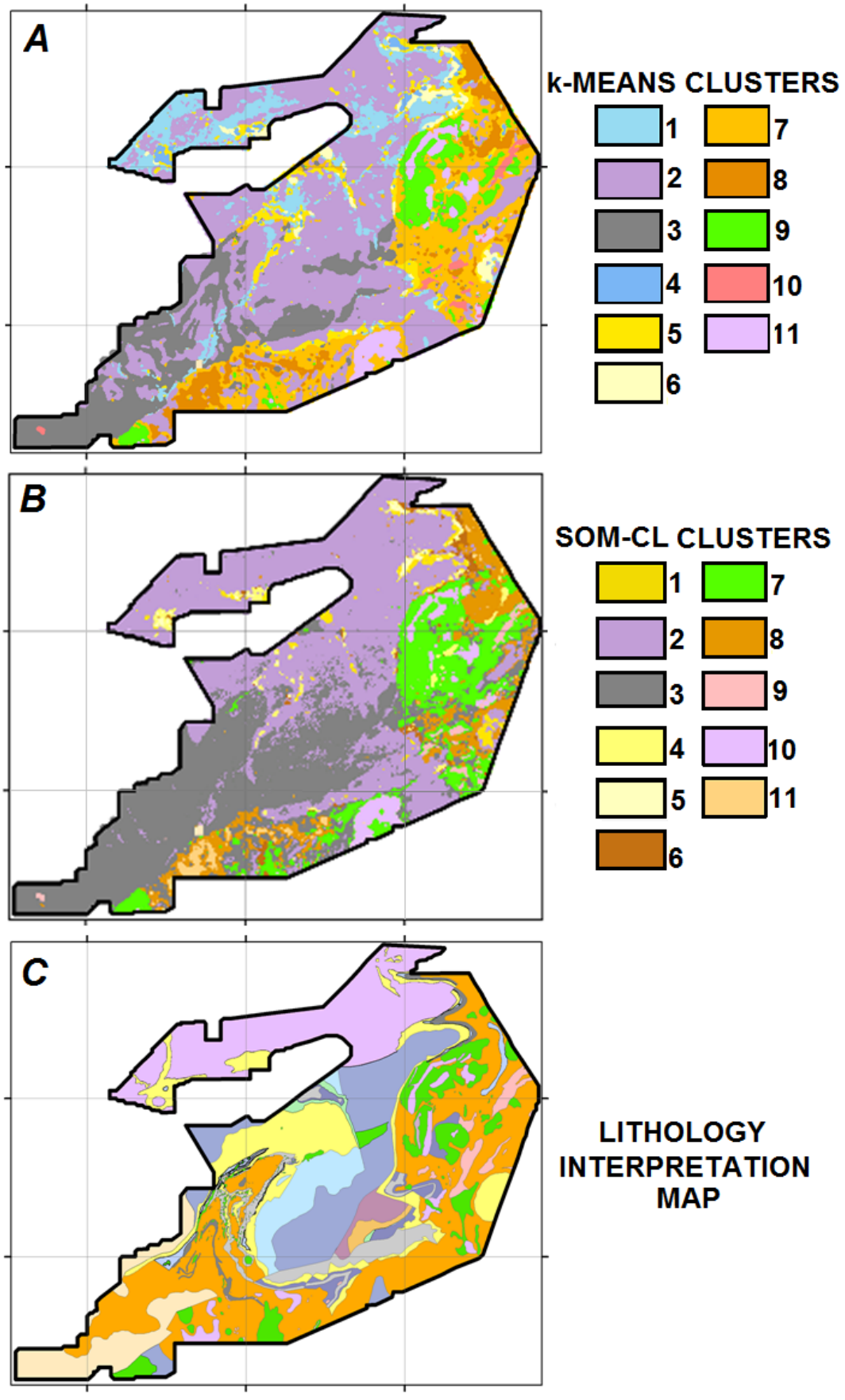


FIGURE 5.14: Comparison of lithology maps. (A) Generated by clustering using k-means and (B) generated by clustering using SOM-CL. (C) The initial interpreted lithology map (Figure 2) is replotted at the same scale to facilitate a visual comparison (C). Clusters are coloured for the best comparison for that clustering output with initial mapped lithology.

and variations at a scale smaller than lithological domains. As such, the absolute magnitude of magnetic response may be diagnostic of lithology at the scale of this investigation, the 1VD is not. This is counter to common use the 1VD as a primary mapping and interpretation tool. In this case, objective ranking would suggest that while the 1VD may be very useful for the mapping of structure, texture, or sub-unit differentiation, it is not diagnostic of lithology.

5.5.2 Classification from outcrop T_a (case study C1 and C2)

Poor sample balance and distribution, in addition to the absence of five lithological classes in outcrop-based T_a, resulted in poor classification results. The complete loss of geometry (Figure 5.8A) reinforces the need to attempt to address class imbalance. In case study C2, results were improved by statistically rebalancing classes by bootstrapping where sample size was inadequate; and randomised decimation where sample size reduction was required. This cannot address the problem of limited outcrop distribution but will correct for the bias introduced in RF due to class imbalance. In this case, while a pixel by pixel accuracy compared to the geological map was still low, correct contact geometries were more closely recovered in the east of the map. Additionally, some classes, namely those with better spatial representation in the training data, were predicted in a more geological reasonable manner (Figure 5.9A, 9B). Care should be taken in rebalancing. Reduction of sample size risks excessive removal of real data, while oversampling preserves real data but introduces a high level of artificial samples (Table 5.2). Caution must be taken when bootstrapping, as this can result in duplicated samples being orders of magnitude more numerous than original, unique samples, producing a tightly defined, over fitted class signal. In such cases, cross validation using training data was misleading (C2, Figure 5.7). RF can produce strong classification results based on over-fitted class T_a, with these results not being indicative of predictive power for new samples. In line with the pragmatic approach taken in these studies, this simple method does not attempt to predict the distribution of sample populations beyond that which was observed. These results therefore could potentially be improved through the use of further strategies for addressing class imbalance if needed for the given exploration goal.

This sample paradigm (C1 and C2) was designed to simulate the state of the project prior to the completion of a robust interpretation map. In this scenario, the extent to which outcrop is representative unknown and explorers will require outputs of RF to assess if or where classification was robust. When classifying new data occurring outside of the spatial range of Ta (outcrop), prediction of the class label of the nearest training data was common. This occurred most notably, in the southwest (Figures 8A, 9A and 9B). These predictions were associated with anomalously low H (Figures 8B, 9C and 9D). We interpret this effect as the being a result of high similarity to a single, most proximal class and low similarity to all other, non-proximal classes. In this case, RF lacks examples of how all but the Roan Group (MSO) classes manifest in the southwest. Contrary to the well documented behaviour of uncertainty calculated from RF class membership probabilities (Kuhn et al., 2018; Kuhn et al., 2016; Cracknell and Reading, 2014; Cracknell, Reading and McNeill, 2014) H and H_{norm} associated with this bulk, incorrect prediction is very low. This anomalous low H, in association with an adjacent class being "extrapolated" away from training data is in fact a key indicator that predictions in that area are incorrect and additionally, indicate that area of the map in question is

distinctly different in data space, to that described by the Ta used. This indicates a spatial transition into an unsampled geological domain but could also occur when presented with rock types not included in the Ta, regardless of spatial range.

5.5.3 Reclassification of geological interpretation map (case study C3)

RF produced a classification output after training on T_a sampled from FQMs most recent geological interpretation map. Overall, the consistency of C3 with the initial geological interpretation map was moderate, at 67.5% (Figure 5.10B). In many cases, classification results were strong, with nine classes achieving greater than 75% consistency with the geological interpretation map. As expected this result is considerably better than the predictions based on limited outcrop (C1, C2) and is consistent with other findings (Kuhn et al., in review, 2018; Cracknell and Reading, 2014) that the results of such RF classification implementations are highly sensitive to an adequate spatial distribution of T_a representative of the range in observed values for a project. A major source of inconsistency with the map is the re-classification of the undifferentiated Kundelugu Group rocks (SOO) into adjacent classes, most notably, the magnetite-altered Kundelungu rocks and the adjacent dolomitic (MCB) and Syenite (ISY) units. It is likely that the original interpretation of the eastern region as undifferentiated Kundelungu rocks is an oversimplification and RF is partitioning rock units within this agglomerate group into correct subdivision, which in turn is supported by a lower H_{norm}.

The geological interpretation map is variably accurate with respect to the real geology of the region as the location and degree of inaccuracies are not quantifiable. As such we have referred to the consistency of RF output with respect to this map, recognising that where inconsistent, it may be the RF prediction, the FQM interpretation map, or both that are incorrect with respect to the real geology. It is a potentially useful insight that the relationship between RF and the starting map is interactive: the interpretation map can be used to validate RF classification, while the RF classification can be used as a form of objective audit of the interpretation map which may demand a small or large scale refinement to the original map. The added benefits of this approach, in addition to the reproducibility of the RF classification are the additional metrics produced by RF. Class membership probabilities (Figure 5.11) can be used to better understand the confidence in prediction of lithology on a per-unit basis. Quantified uncertainty, in the form of H (Figure 5.12A) and H_{norm} (Figure 5.12B) relate to the difficulty of assigning a correct lithology to a given sample and the associated data. It is reasonable to assume that this ambiguity, a function of the expression of the data at a given location, influences any other manual attempts at classification using these data and thus H and H_{norm} facilitate review not only of the RF classification, but also other manual mapping efforts. H defines areas of geological complexity, frequently tracking lithological boundaries (Figure 5.12A). In this case, areas of high H, related to those with the greatest number of possible lithologies present, include most notably, the geologically and structurally complex fold hinge in the central-west and a large region of the central-south (Figure 5.12A). H_{norm} displays the uncertainty of each sample, relative to its own minima and maxima, independent of number of possible classes. This can be seen in comparing Figure 5.12B, where a larger number of pixels exhibit high H_{norm} (warm colours), with 12A where the number of pixels with high H is lower, de-emphasising areas with fewer classes. High H_{norm} is correlated with a higher probability of incorrect classification. Of further

interest are regions where RF has made classifications with low associated H_{norm} that are inconsistent with the starting interpretation map. This may indicate regions where RF has made a correct prediction against an incorrect starting map. A notable example is the partitioning of undifferentiated Kundelungu rocks (SOO) into the magnetite-altered Kundelungu Rocks class (AOO) described above. This more extensive domain of AOO class, identified by RF is not apparent in the RTP data. As the number of classes incorporated in this study was higher than was the case for C1 and C2, the absolute range of H is not comparable across the 3 studies.

5.5.4 Mapping via clustering (case study C4)

Both k-means and SOM-CL, when unconstrained by the number of clusters, converge on clusters that can easily be mapped to major tectonic domains. When constrained to a minimum reasonable number of clusters, based on outcrop mapping, both clustering methods converged on an optimal number of 11 clusters. Clusters were produced that showed a strong spatial resemblance to lithology (Figure 5.14). K-means clusters showed a stronger spatial correlation with interpreted geology, however, with apparent sensitivity to drainage patterns. SOM-CL clusters by comparison were less sensitive to drainage patterns and performed well in recognising clusters spatially congruent with the Kundelungu Group sub units while grouping the region associated with gabbros with much of the neighbouring Kundelungu Group. Both methods reveal a large loosely semi-circular cluster in the central south of the project This cluster shows spatial congruency with syenite rocks in the central-east (Figure 5.14: k-means cluster 11 and SOM-CL cluster 10).

K-means is relatively easy to implement and understand conceptually. Additionally, k-means is fast, with clustering results for this study produced in minutes, using a high end (at the time of this study) but standard production desktop PC. This is an important factor for uptake by exploration teams as there is no requirement for specialised computing skills. Speed of analysis facilitates iteration, experimentation and modulation of input variables. SOM is a more sophisticated algorithm and the additional steps associated with SOM and hierarchical clustering, as with SOM-CL used in this study adds further demands on the user. The algorithm is potentially capable of identifying more complex groupings in data than k-means. The caveat is that SOM-CL requires significantly more sophisticated tuning which in turn requires some degree of specialist knowledge for robust implementation. Additionally, SOM run times are significantly longer than k-means and do not lend well to repeat experimentation. It is worth considering that geoscientific data in the 2D map space does not exhibit the level of complex, non-convex datasets seen in other computing fields. With that in mind, we assert that k-means is an adequate starting point for a 2D mapping problem and may perform as well, or better than more sophisticated algorithms. SOM-CL also produced excellent results in this study and the additional flexibility in tuning and production of validation metrics make it a valuable addition to the toolbox and a useful option for cases where more complex data are encountered, or a more comprehensive understanding of dataset topology is desired.

5.6 Conclusions

Our testing of Random Forests classification and clustering methods using the CACB Trident dataset identified a number of machine learning usage strategies likely to be of value to create/improve the working lithology map at both early and mature stages of mineral exploration. Dataset ranking, and prioritisation should be undertaken. The rankings produced by RF formed an important part of the classification process and provide information that assists in optimising clustering results. They also serve as a prompt to assist conventional geological interrogation.

Machine learning algorithm usage strategies that we found to be important in scenarios replicating early stages of geological exploration ensure that a meaningful lithological map is produced and that a quantitative appraisal of inaccuracy may be made. RF classification using a limited training dataset, naively sampled from raw outcrop information, results in low classification accuracy. In such circumstances, RF results are not meaningful. Balancing class sample size produces optimal results from a restricted training dataset, better predicting some classes and improving recovery of mapped geometries while noting that high cross validation accuracy is not indicative of predictive power for new samples. The spatial extent of the training data needs to be considered to avoid the over-extended prediction of a boundary proximal class. Such boundary proximal class predictions, away from T_a and coupled with low H, can be interpreted as a warning sign that predicted classes are encroaching into regions comprising lithologies not represented by T_a.

Machine learning algorithm strategies appropriate for scenarios replicating more mature stages of exploration were demonstrated with the classification of lithology from a training sample comprising Ta from an existing interpretation map. The use of RF in such in data-rich exploration settings is very valuable, leveraging the additional information available, in producing a more accurate and insightful prediction. Using RF at this stage fulfils two important functions: firstly, as a means of performing an objective audit of the starting map; and secondly, as a basis of refining the initial product. H, H_{norm} and class membership probabilities can be used to evaluate RF outputs or better understand the uncertainty associated with both the pre-existing geology map and the refined map produced through the RF prediction.

Clustering is a further tool that may be of utility in lithological mapping. Both k-means (and SOM) produced results showing spatial congruency with mapped lithologies, providing a powerful first pass mapping tool without the need for a Ta. In this study clustering, k-means in particular, produced a map, in the absence of geological constraint, which allocated clusters with close spatial affinity for the position of mapped lithologies as they are currently understood by FQM. This suggests that clusters are responding to lithology above other effects. Alternatively, these methods could be used to appraise, validate or refine an existing map. Geological domain knowledge may then be added to interrogate clusters and assess if/how they relate to lithology, alteration or other geological processes.

5.7 Acknowledgements

We would like to thank First Quantum Minerals Ltd. for permission to access data. We thank Chris Wijns and Tim Ireland for support and discussion regarding data and results. Stephen Kuhn is supported by a Tasmanian Graduate Research Scholarship (TGRS) from the University of Tasmania. This research was conducted as part of the ARC Industrial Transformation Research Hub for Transforming the Mining Value Chain (project number IH130200004) at the Centre of Excellence in Ore Deposits, University of Tasmania. The views expressed herein are those of the authors and are not necessarily those of the Australian Research Council. We used the Orange software package (Demsar et al., 2013) for RF classification and k-means clustering, and the R package: Kohonen (Wehrens and Buydens, 2007) for SOM. Pre-processing, interpolation and plotting were performed using Geosoft Oasis Montaj and ESRI ArcGIS.

References

Arthur, D. and Vassilvitskii, S. (2006). *k-means++: The Advantages of Careful Seeding: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, New Orleans, Louisiana.

Bank of Zambia Annual Report (2016). `http://www.boz.zm/annual-reports.htm`. Accessed October 2017.

Beck, F., Burch, M., Munz, T., Silvestro, L. and Weiskopf, D. (2014). Generalized pythagoras trees for visualizing hierarchies: 9th international conference on information visualisation theory and applications.

Berlein, F., Fraser, S., Brown, W. and Lees, T. (2014). Advanced methodologies for the analysis of databases of mineral deposits and major faults, *Australian Journal of Earth Sciences* **55**: 79–99.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

Breiman, L. (2001). Random forests, *Machine Learning* **45**: 5–32.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth Brooks/Cole Statistics/Probability Series, Wadsworth International Group.

Capistrant, P., Hitzman, M., Kelly, N., Kuiper, Y., Wood, D., Williams, G., Zimba, M., Jack, D. and Stein, H. (2015). Geology of the Enterprise hydrothermal nickel deposit, *Economic Geology* **110**(1): 9–38.

Cracknell, M. (2014). *Machine Learning for geological mapping: Algorithms and applications.*, University of Tasmania: University of Tasmania.

Cracknell, M. and Reading, A. (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): 113 – 126.

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

Cracknell, M., Reading, A. and McNeill, A. (2014). Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests^{TM} and Self-Organising Maps, *Australian Journal of Earth Sciences* **61**: 287–304.

Davies, D. and Bouldin, D. (1979). A cluster separation measure, *IEEE Transactions on Pattern Analysis Machine Intelligence* **1**(2): 224.

Defays, D. (1977). An efficient algorithm for a complete link method, *The Computer Journal* **20, 4**: 364–366.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. and Zupan, B. (2013). Orange: Data mining toolbox in python, *Journal of Machine Learning Research* **14**: 2349–2353.

Fraser, S. and Dickson, B. (2008). A new method for data integration and integrated data interpretation: Self-Organising Maps.

Guyon, I. (2008). Practical feature selection: from correlation to causality, *in* F. Fogelman-Soulié, D. Perrotta, J. Piskorski and R. Steinberger (eds), *Mining Massive Data Sets for Security – Advances in Data Mining, Search, Social Networks and Text Mining, and their Applications to Security, 19*, IOS Press, Amsterdam, p. 27–43.

Harris, J. and Grunsky, E. (2015). Predictive lithological mapping of Canada's north using Random Forests classification applied to geophysical and geochemical data, *Computers and Geosciences* **80**: 9–25.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics* **v. 43**: 56–69.

Kohonen, T. (2001). *Self-organizing maps*, Springer series in information sciences, 30, Springer-Verlag, Berlin.

Kuhn, S., Cracknell, M. and Reading, A. (2008). Lithological mapping via Random Forests: Information entropy as a proxy for inaccuracy, *25th International Geophysical Conference and Exhibition, ASEG, Extended Abstracts* p. 1–4.

Kuhn, S., Cracknell, M. and Reading, A. (2018). Lithological mapping using Random Forests applied to geophysical and remote sensing data: a demonstration study from the Eastern Goldfields of Australia, *Geophysics* **83**: B183–B193.

Liaw, A. and Wiener, M. (2002). Classification and regression by Random Forests, *R news* **2**: 18–22.

Lloyd, S. (1957). Least squares quantization in pcm: Technical note, Bell Laboratories, *Published in IEEE Transactions on Information Theory* **28**(2): 129.

MacLean, W. and Barrett, T. (1993). Lithogeochemical techniques using immobile elements, *Journal of Geochemical Exploration* **48**(2): 109–133.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*, University of California Press, Berkeley.

Mining for Zambia, A concentrated mining sector (2016). Accessed October 15, 2017. **URL:** *https://miningforzambia.com/a-concentrated-mining-sector/*

Pearce, J. and Norry, M. (1979). Petrogenetic implications of Ti, Zr, Y, and Nb variations in volcanic rocks, *Contributions to Mineralogy and Petrology* **69**(1): 33–47.

Penn, B. (2005). Using Self-Organizing Maps to visualize high-dimensional data, *Computers and Geosciences* **31**(5): 531–544.

Rodriguez-Galiano, V., Chica-Olmo, M. and Chica-Rivas, M. (2014). Predictive modelling of gold potential with the integration of multisource information based on Random Forest: a case study on the Rodalquilar area, *Journal of Geographical Information Science* **28**(7): 1336–1354.

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

Selley, D., Broughton, D., Scott, R., Hitzman, M., Bull, S., Large, R., McGoldrick, P., Croaker, M. and Pollington, Y. (2005). A new look at the geology of the Zambian Copperbelt, Vol. Hundredth Anniversary Volume, pp. 965–1000.

Shannon, C. (1948). A mathematical theory of communication, *Bell Systems Technical Journal* **27**: 379–423.

Ultsch, A. and Vetter, C. (1994). *Self-organising Feature Maps versus Statistical Clustering A*, Deptartment of Mathematics and Computer Science, University of Marburg, Benchmark.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer-Verlag New York, Inc.

Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley Sons Inc.

Waske, B., Benediktsson, J., Árnason, K. and Sveinsson, J. (2009). Mapping of hyperspectral aviris data using machine-learning algorithms, *Canadian Journal of Remote Sensing* **35**: 106–116.

Wehrens, R. and Buydens, L. (2007). Self- and super-organising maps in r: the Kohonen package, *Journal of Statistical Software* **21**: 1–19.

Wellmann, J. and Regenauer-Lieb, K. (2012). Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models, *Tectonophysics* **526–529**: 207–216.

Chapter 6

Synthesis

The research presented in this thesis has expanded knowledge through demonstration studies of the situation specific application of machine learning (ML) for lithological map production and refinement in a mineral exploration context. The overarching objective of this thesis was to progress the use of using ML for this purpose, and the insights from the core science chapters are now synthesised in this final chapter.

6.1 An Expanded Set of Demonstration Studies

Chapters 3, 4 and 5 each defined a case study, or studies, in a different geological setting and target deposit style. These studies all take, quite deliberately, a pragmatic approach with regards to available data and position in a mineral explorer's workflow, leading to demonstrable practical applications for explorers looking to incorporate ML into their geological mapping efforts. The studies focus on approaches that can be replicated by most geoscientists, without requirements for specialist programming and/or GIS skillsets, identifying where such methods work effectively, where they do not, and where improvements might come from more sophisticated methods.

Importantly, the research demonstrates effective usage of supervised and unsupervised ML at various stages in the maturity of an exploration project. Thus, the data available at that point in time is utilised appropriately to inform the subsequent exploration stage.

6.1.1 Random Forests lithology classification studies

This thesis, and the papers contained herein, extends the usage of Random Forests (RF) for surficial solid geology mapping in a mineral exploration setting over a wide sample of geological domains and expected commodities. Cracknell et al. (2014) demonstrated the efficacy of the RF for lithological map refinement, making use of a small volume of spatially well distributed training data in western Tasmania, a region of significant and well established volcanic hosted massive sulphide (VHMS) style mineral endowment. The three technical papers comprising this thesis extend widely the usage of this method in several ways. Firstly, these new studies extend to the exploration of a wider range of deposit styles: porphyry / epithermal systems (Cu, Au, Mo, Ag; Chapter 4, Kliyul, British Columbia, Canada), orogenic gold (Chapter 3, Heron South, Eastern Goldfields, Western Australia) and sediment hosted copper, cobalt and nickel systems (Chapter 5, Trident, Central African Copper Belt, Zambia). While each study represents a

sample of one, works comprising this thesis strongly suggests that the methods applied can be effective in the settings hosting a large proportion of the major commodities.

Each study comprising this thesis encountered unique challenges specific to the geological setting in which they were situated. This included flat lying areas under significant cover, both transported (Chapter 3) and residual (Chapter 5); areas of significant topography (Chapter 4); and terrains ranging in age from Archaean (Chapter 3) to Proterozoic (Chapter 5) to Mesozoic (Chapter 4). Additionally, the available data for each project reflected in part both the geological setting and the stage of maturity in the exploration pipeline of each. Chapter 3 represents an early stage conceptual target area, reflected in the available data being limited to remote sensing and potential field geophysical data. Chapter 5 (case study 1) presents in some ways a similar set of conditions at a relatively early exploration stage, however due to the existence of active mining operations in the area and the presence of mostly residual soils; soil geochemical data were available, adding significantly to the information available for all forms mapping by RF. Likewise, Chapter 4 presents a study with good ground based geochemical sampling and a significant spread of outcropping geology, allowing mapping to focus on more discrete units. Together, these studies demonstrate that the ML methods described can be effective at most stages likely to be encountered by mineral explorers. Perhaps the most important consideration is the availability of spatially well distributed training data, which, when available, delivered consistently strong results. Chapter 5 describes a scenario in which training data are restricted to spatially discrete and not well representative regions of the study area. Further studies of this type of situation would be useful to build experience in the production of new / first pass geological maps in the real, non-idealised outcrop faced when exploring new ground, over which explorers have no control.

6.1.2 Objective audit of a pre-existing lithological map

A key aim of this research was to demonstrate the efficacy of RF for the refinement of an existing map. This can be thought of as two steps, firstly, the generation of a new map, and secondly, an objective appraisal (or audit) of the previous version. This was accomplished in core research Chapters 3 and 5. In each case, RF, using a sparse training sample derived from the existing geological map, was able to achieve a high level of consistency with existing geological mapping. Chapter 3 represents an early stage exploration project comprising airborne and remotely sensed reconnaissance level data and geological mapping produced through desktop interpretation. Chapter 5 includes (though with inconsistent distribution across the project) the benefit of geological mapping and comprehensive soil geochemistry. In these examples, training data were taken via a spatially balanced random sample from an existing geological map. Lithological mapping using RF offers, in a form that is repeatable and with quantifiable uncertainty, a means of objectively auditing an existing an existing map. This is achieved by taking a small, spatially balanced sample with equal representation from each mapped lithology: in the order of 2% of the total area. This allows an accurate class expression to be defined by RF while allowing the majority of the map the freedom to be reclassified. In this manner, RF can identify inconsistencies where a mapped unit does not coincide, in data space, with an explorer's own mapping (or unit

definitions) throughout the sampled locations in the map. More generally, this approach draws attention to where lithology prediction is well supported by the available data and where it was not. This may also point to inconsistencies in geophysical and geochemical character within a broader lithological classification, such as was seen in Chapter 5 (case study 3), where one oversimplified stratigraphic grouping was split into more specific subunits. While these had been inferred in absence of outcrop, evaluation of the data using RF clearly confirmed that these subunits, identified elsewhere in mapping, were present.

6.1.3 Refinement of a pre-existing lithological map

Any observed inconsistency between an original geological map and the RF produced product using that map as a basis for training data must be reconciled either through the identification of incorrect predictions by RF or incorrect, or imprecise mapping in the original. This in turn requires modifications to the existing map or need for better understanding of the variable expression of data across a lithological unit that was indeed mapped correctly. This can be supported or assisted by the quantitative metrics such as class membership probabilities (CMP) and uncertainty, made possible through the ML approach. This was well demonstrated in Chapter 3, where RF predicted a transition from mafic stratigraphy to granite several hundred metres further to the west than the original mapping suggested. The behaviour of information entropy (H; Shannon, 1948), increasing, approaching a maxima over a lithological boundary, while remaining low over the previously mapped boundary, supports the RF prediction and indicated that the prior mapping requires revision.

6.1.4 Clustering

Clustering was a relatively minor point of focus in this research, although it provides a simple and intuitive reference method of potential utility. It was trialled as an alternative to classification for the task of map production. In Chapter 5, case study 4, both k-means and Self-Organising Maps (SOM) clustering algorithms were used. Firstly, they were applied to all datasets, and again applied only to those datasets identified as non-redundant according to the RF variable importance ranking, with the results of the latter outperforming the former. Both methods were subjected to some manual guidance via use of *a priori* geological information. For k-means, this took the form of restricting the number of clusters to a reasonable range, based on the number of rock types observed in the geology and in the case of SOM, complete-linkage hierarchical clustering was used to group nodes into amalgamated clusters of a geologically meaningful number (in both cases, from 10 to 20). Interestingly, prior to the application of such constraints, both algorithms optimised at a low number of clusters (3 and 5 for k-means and SOM respectively) which correlated well with major tectonic domains. Once limited to a number of clusters feasible to represent the surficial geology at the project scale, both algorithms converged at 11 clusters. In this case, k-means returned the superior map (as compared with the most well-developed company geological map and that produced by RF in Chapter 5 (case study 3)), better preserving the structural and geological complexity of the project. While it is entirely plausible that further tuning of the SOM parameters would with time, produce a superior map, the fact

that a simple algorithm such as k-means was able to perform well is an important conclusion to take forward. This is of importance for wider deployment of clustering algorithms in the geosciences by practitioners without expertise in ML algorithms or computer sciences.

6.2 Prediction evaluation and metrics for knowledge generation

A key component in the usage of ML for mineral exploration is the ability to produce repeatable, quantified, and hence, objectively assessable outputs. The ability to quantify, express and visualise the relationship between possible classes and an overall uncertainty is a distinct point of difference between ML based map prediction and conventional manual interpretation by geoscientists. The complex nature of results may be captured, as is the potential for inaccuracy both for any given sample and across a project area. This prediction evaluation is distinct from many more mainstream fields of ML, for example robotics, voice or character recognition or virtual personal assistants. Challenges intrinsic to mineral exploration frequently result in classification inaccuracies: limited and poor training data, variable sampling, under sampling, non-uniqueness of the property, and position of sources of potential field signals. So far as an existing geological mapping can be used as the basis for comparison, the success rate is of the order of 60-80% (as demonstrated in Chapters 3, 4 and 5 of this thesis). Given the relatively large proportion of error still present even under the best of normally encountered circumstances, the ability to assess the nature and possibly cause of misclassification is critical. This thesis presents several metrics for variable assessment and ranking, competing class probabilities and quantified uncertainty.

6.2.1 Variable ranking and reduction

In Chapter 5, it is demonstrated that through variable ranking and reduction, utilising the objective method embedded within RF, results of clustering are improved. This is an important benefit to the method as it allows the generation of more accurate classification results, which can both pre-empt and assist more detailed and time consuming geological, geophysical, or geochemical interpretation of the same datasets. Furthermore, the ranking, reduction, and improved clustering and classification results may be used to assist in building understanding for a new project, potentially saving time both time and money while being objective, repeatable and available to geologists regardless of experience and skill level. The studies in this thesis have shown that many of the variables ranked highly correspond with those well understood in conventional mapping. This behaviour is to be expected to some extent given the RF has been tasked with predicting lithology. This provides valuable confidence for geologists utilising the method. Variable ranking, as shown in all case studies comprising this thesis routinely includes datasets which may not be obvious to geologists at first glance, or, would have been a lower priority in order of investigation, potentially expediting understanding of a project. A further benefit of this process, shown in Chapters 3, 4 and 5, is the reduction of dimensionality down to those datasets necessary to make accurate predictions. This facilitates a more manageable

interpretation of results by the end user. The case studies comprising this thesis, along with work by Cracknell et al. (2014) have shown that RF, in this context spanning a range of ore deposit types, has consistently required between 8 and 15 variables to produce an optimal result; while often starting with in excess of 50 to 100 variables. Considering 8 to 15 variables simultaneously yields benefits over traditional interpretation by a geologist (utilising comparisons of 2 or 3 variables, or a mild dimensionality reduction through ratios) yet is still a small enough number of datasets for a geologist to meaningfully interpret following the ML analysis.

6.2.2 Uncertainty

A key benefit for explorers using a ML approach is the ability to quantify uncertainty. RF's ensemble approach: constructing many uncorrelated classification trees allows the user to investigate the distribution of classification results. The proportion of trees voting for a given class, of the total number of trees comprising the RF are described as CMPs. In the studies comprising this thesis, H is the chosen metric used as a proxy for expressing uncertainty. H is described in detail in Chapter 2, and, where this metric is used in Chapters 3, 4 and 5. H is a measure of the of the disorder, or lack thereof as expressed by competing CMP. An important criterion for the use of H as the preferred metric for expressing uncertainty for the research comprising this thesis is the preservation of monotonicity, with the addition of terms, i.e. possible classes, contributing to a higher possible H. This means that H will increase in response to uncertainty as well as complexity. Both of which can be well expressed by assessing and visualising H and normalised H. This allows for the identification for areas of greatest lithological, and in many cases, by inference, structural, complexity. Furthermore, H can be normalised (H_{norm}) on a per-instance basis, by number of classes present to express to what extent each instance approaches its own maximal possible disorder. Prediction uncertainty, in the form of H, is an expression of the internal consistency of the RF prediction and is not, nor is intended, as a direct measure of the congruency between the RF prediction and reality. In this thesis, however, it is demonstrated that there is good correlation between high H_{norm} and where predictions were inconsistent with observed or mapped geology (noting that such maps may themselves contain inaccuracies). H_{norm} grouped by accurate and inaccurate prediction are shown to comprise statistically distinct, though overlapping populations (Kuhn et al., 2018; Kuhn et al., 2016). Looking at H_{norm} where predictions can be validated (i.e. areas of known geology within the study area) the statistical distributions of both populations should be displayed and analysed. This allows an explorer to tune a cut off value that corresponds to their particular risk tolerance. For example, by choosing a higher cut off value, a more complete map can be retained, albeit with an increased rate of misclassified samples being included, or, through use of a lower cut off value, the number of misclassified samples can be minimised, but at the expense of some proportion of correctly classified data also being excluded. The full distribution of H_{norm} for accurately vs inaccurately classified samples (where knowable) should always be analysed in detail prior to using H_{norm} for the purpose of excluding a classification that is likely to be incorrect.

Where a correlation between H_{norm} and prediction inaccuracy can be clearly demonstrated, the observed relationship between H and lithological complexity is of a more anecdotal nature. Cracknell et al. (2013) demonstrated a relationship

between high uncertainty and proximity to lithological contacts, a form of lithological complexity, as predicted by RF, while Wellmann (2012) showed a clear relationship between high H and the number of possible lithologies present at a location. The studies comprising this thesis are consistent with these findings, showing that H increases in areas where an increasing number of lithologies are present. This is contingent on a simple definition of lithological complexity as being more lythotypes possibly present. Any further investigation of lithological complexity in a spatial sense requires a specific and somewhat subjective definition of geological and structural complexity and a rigorous assessment of the behaviour of H against a number of such definitions would be a valuable and productive avenue of future research.

6.2.3 Assessing spatial limits of predictive capability

A key component in the application of ML in the geosciences are the considerations regarding the spatial distributions of both training data and extents to which prediction is applied. Through comparisons of map audit and refinement studies in Chapters 3, 4, and 5 where data were spatially well distributed and balanced; against Chapter 5's cases 1 and 2, and synthetic examples presented in Cracknell & Reading (2013), it is evident that spatially well distributed data are a key factor in the ability to make accurate predictions. This is intrinsically linked to the likelihood of capturing i) all geological units, and ii) the full range of expressions of each unit, within a project area increasing as more of the area is sampled. In Chapter 4 data are taken from the class labels at soil/rock chip sample locations collected on a nominally regular grid. In all cases, training data are spatially well distributed across the project areas. This contrasts with additional case studies in Chapter 5 where spatially discrete geological observations, in the form of outcrop, are taken as training data. While measures were taken to add value and improve results; classification performance was poor as compared to where a spatially more comprehensive sample was used. While the relationship between an equal representation of all classes, from a statistical standpoint, and an even spatial distribution of training data may vary, these properties are linked in practice. It is apparent that a spatially representative spread of training data contributes to a better classification result. This need not be the case as there will be examples where limited outcrop may represent the full expression of geology in an area. This is, however, not easily predictable and will vary widely on a case by case basis and must be assessed thoroughly with best available information at the outset of a study and considered carefully when assessing results.

The spatial resolution of classification or clustering exercises are inherently linked to that of the input data. This introduces a number of challenges with regards to how data are interpolated and ultimately how data acquired at various resolutions are amalgamated. Potential field geophysical data were gridded in accordance with current industry best practice methods, namely a combination of minimum curvature and bi-directional splines at a grid cell size equal to a quarter to a fifth of the survey line spacing. Geochemical data were gridded at half to a third of sample spacing tightly controlling the level of interpolation allowed taking an approach more akin to a nearest neighbour. Remote sensing, being of a pervasive nature already, were treated as-is. In order to perform a pixel/instance based classification, all datasets must be sampled on a consistent grid. The variability between sample

intervals of incorporated data will inevitably result in some being undersampled while others oversampled. In the studies comprising this thesis, an approach towards oversampling was taken. This allowed classification to benefit from the resolution permitted by higher resolution datasets. It is interpreted that by incorporated upsampled lower resolution datasets that overall class accuracy is improved at the cost of minor degreadation of spatial accuracy, particularly near class boundaries. lower spatial resolution does not appear to have a deleterious effect on variable importance, as evidenced by the high ranking of geochemical data (lower spatial resolution) in chapters 4 and 5. Further study however, on the effects of isolating the inclusion or exclusion; or modulation of resolution, of select datasets on variable importance would be beneficial.

With regard to the extent of the domain to which a classification can be applied, the presence of lithologies not present in training data must produce an erroneous result when encountered by a classifier. Unfortunately, in a genuine exploration context where the rocks are unknown, the explorer will not know of rocks they are yet to see but will also be unaware of where such a misclassification has occurred. This is inherently unknowable at the outset of a study without a degree of explicit prior information, such as reliable prior geological mapping, conclusive evidence of a domain change in geophysical data, indications from exploratory data analysis, including clustering methods, or other lines of evidence. In Chapter 5 a result is seen associated with incorrect classification, whereby RF begins to predict the most boundary proximal class, associated with low H, away from known observations. A departure from the domain in which the training data are representative of the rocks being classified is thereby suspected. The most probable cause of this result is that broad similarities in geophysical data, based on proximity between the nearest training sample and the "out-of domain" sample are overriding other similarities. Nevertheless, this is a key problem that warrants further investigation. It may be unsolvable, in that rocks from an adjacent area may be identical in property space to those within the predictive domain. Thus, attention should be directed to the best possible definition of the study domain while remaining alert to any possible indications of departure from the domain.

6.2.4 Assessment of class membership probabilities

The ability to quantify uncertainty in classification is a key benefit of classification using RF, and indeed ML approaches more generally. The use of qualitative and quantitative interpretation of CMPs in this regard can be highly informative. Chapter 4 focusses on lithological classification in a porphyry exploration context. In this case, discrete intrusions can be identified through an increase in CMP, relative to a background rate, regardless of whether they were correctly classified in final classification output or misclassified as their more spatially extensive surrounding host rocks. This is a very valuable finding and allows the prediction of source / host rocks for porphyry mineralisation amongst complex volcanic, volcaniclastic and volcano-sedimentary rock packages with overlapping class signatures. In such instances, an accurate geological map of a project might be difficult to produce, but a method to point directly to potential source intrusions is perhaps more valuable still. This study (Chapter 4) demonstrates the efficacy of using elevated CMPs at an early stage of the exploration cycle and could guide targeting prior to mobilisation of a drill rig.

Further to examining individual class predictions and the outright result, where misprediction does occur, assessments can be made of which classes are commonly confused and the rate at which misclassification occurs. Using confusion matrices compiled during cross validation on training data, one can analyse where, i.e. between which classes, misclassification is likely to occur. It is then a reasonable assumption that a similar misclassification between those classes is more likely when applying that classifier to the remainder of a project area. Furthermore, this information can be valuable independently of classification in highlighting geochemical and petrophysical similarities between lithological and/or stratigraphic units that may not yet be well understood at the time of classification.

6.3 Lithological Mapping Aided by Machine Learning in Mineral Exploration

Each research chapter comprising this thesis is representative of a specific stage of mineral exploration. Collectively, these studies demonstrate that RF lithological classification can be of value at any stage of a mineral exploration project.

6.3.1 Adding value through machine learning at different stages of project maturity

As demonstrated in Chapter 3, the method can be used to refine an interpreted geological map using geophysical data. Chapters 4 and 5 each comprise a study using direct geological observations as training data for and RF classification and included the use of soil / rock geochemistry, representing a more advanced exploration stage than was the case in Chapter 3. In each instance the method was effective in adding value to the project. In Chapter 4, this was achieved directly through the identification of intrusive rock units, providing a direct target for drilling or further mapping or sampling. In the equivalent case studies in Chapter 5 (case studies 1 and 2) a poor distribution of outcrop resulted in low accuracy in classification however, the knowledge gained from variable ranking could provide a valuable head start on geochemical interpretation and the pre-selection of those variable deemed non-redundant by RF ranking, to the exclusion of all others, improved clustering results.

The key difference between these studies was the spatial extent of training data relative to the full extent of the project area. Unsurprisingly then, a larger range of outcrop spread more widely over a project will contribute to a better result, though it is entirely possible in some circumstances that the full expression of the geology of a project area could be obtained via sampling of a spatially limited region. If certain rock types are unrepresented in the training data, they may be excluded from the training data by the simple fact that they are not present in outcrop. Alternatively; geologically, petrophysically and geochemically similar rock types may exhibit significantly different expressions in geophysical and to a lesser extent geochemical datasets, due to regional trends not related to geology at the scale of mapping. One way to explore this latter possibility is through modelling of potential field data at the scale of the project. This approach could yield meaningful improvements if the model is accurate or have a deleterious effect if it is not. This

approach was not pursued in this research due to the stated goal of objectivity in classification and the inherent subjectivity in the execution of modelling. Nevertheless, it is recognised that combined approaches (ML and traditional geophysics) could be of utility where confidence can be had in geophysical modelling.

As discussed in Section 6.1, using RF to perform an objective audit on an existing geological map is a valuable process at any stage of exploration. By using a small proportion of samples (1-5% as demonstrated across Chapters 3 through 5) as training data, adequate freedom is allowed over the majority of a project area permitting deviation from the starting map where patterns in the training data indicate. Not only does this present the opportunity to refine the geological map; this facilitates the provision of variable ranking, CMPs and H, to supplement existing geological understanding and further highlight any limitations in current mapping products.

6.3.2 Practical considerations for mineral exploration

The research methodologies used in this study were progressed with the non-expert in ML in mind. The bulk of data preparation, pre-processing, data visualisation and assessment can be performed using industry standard software packages. Only the execution of a chosen ML approach, requires specialist software or code with, Python - Scikitlearn, used directly and as implemented in Biolabs' Orange package (Demsar, et al., 2013) and R used in this research. The procedures used are repeatable, building upon the workflow outlined in Cracknell, et al. (2014) to avoid known pitfalls such as not testing for conditional independence of training data, class size imbalance, poor spatial representation of training data, improper data scaling or inclusion of an excessive number of redundant variables.

Where these aspects of input data character could not be avoided, strategies to identify potential errors, and if possible, attain new insight, were investigated. This was most prevalent when restricting training data to outcrop, as seen in Chapter 5 (case study 1), which resulted in both an imbalance in training class size representation in training data. In this research, a simple strategy of bootstrapping and decimation was used, increasing or reducing the number of samples restricted to actual observations, in order to address class size imbalance. It was apparent that training data present at outcrop did not provide enough information to determine an accurate probability density function for more sophisticated imputation and avoid the potential for over-fitting bootstrapping can cause. In this case, the simplest means of addressing class imbalance without unduly throwing away real data was used.

It should be noted that where data of sufficient volume, representation and quality exist, a number of additional approaches can be investigated to extract additional value. These include but are not limited to more complete strategies for imputation such as imputation from variables probability density function via kernel density estimation or similar and more sophisticated strategies for analysing compositional data, such as generating independent orthogonal variables (principle component analysis for example). It is recommended that where practical, and in line with the goals and time frame of the exploration campaign, such investigations be attempted

however they fell outside of the aims of this body of research. The reader is directed to Hood (2019; completed in parallel to work comprising this thesis), Hood et al. (2018), Grunsky & Kjarsgaard (2015) and Mueller & Grunksy (2016); and references therein for a more detailed assessment of these methods.

A summary of the key challenges to the implementation of ML for lithological mapping in mineral exploration is provided in Table 6.1 together with a set of matching recommendations, developed through this research, for addressing each in an accurate, practical and productive manner.

TABLE 6.1: A summary of exploration challenges identified through this research, and recommendations for addressing those challenges through ML approaches.

Exploration Challenge	ML Approach	Recommendations
Lithology prediction in absence of training data	Unsupervised classification / clustering	• If available, cluster using variables as identified through objective ranking (e.g. RF) or as identified through *a priori* geological information as being good predictors of lithology • Ensure all variables are equally scaled to avoid distortion of distances in data space: geophysical and geochemical data whose raw values can be orders of magnitude apart • Let the complexity of the problem dictate the choice of clustering approach. Use a simple clustering algorithm first. If this is inadequate move to a more sophisticated approach • Numerical scoring of cluster separation will show multiple peaks, the most prominent of which is likely to respond to large, first order effects such as major tectonic domains or cover/no cover. Peaks at higher no. of variables may correspond to lithology
Lithology prediction with limited training data for supervised classification	Numerically balance classes and attempt supervised classification	• Use a combination of bootstrapping / imputation and decimation. Balance to avoid throwing away real data vs introducing an excessive number of artificial data • Additional focussed mapping / ground truthing recommended in areas of higher H / H_{norm} where geology is likely to be complex and predictions incorrect
Identification and prioritisation of important datasets	Dataset ranking and reduction via RF or other statistical measures	• Reduce the number of datasets to only non-redundant variables for faster processing and ease of interpretation. • Assess non redundant variables with respect to domain knowledge for sensibility, e.g. large ionic radii / immobile trace elements should factor into lithological discrimination. Look for additional, unexpected predictive variables • Use non redundant variables in place of all variables to improve further classification or clustering results
Audit (± revision) of an existing geological map/interpretation	RF classification trained on spatially and numerically balanced random sample from existing map	• A sample size of ≈ 2% of map area can provide adequate samples for a training set while allowing sufficient freedom form RF to reclassify map • Use H_{norm} to assess likelihood of pre-existing map or RF map being correct. High values of H likely to be incorrect at thresholds guided by RF used in cross validation / training data. H_{norm} indicates likely proximity to geological complexity such as class boundary (contact) or structurally complex area
Identifying subtle / spatially discrete (relative to data spacing) lithological units	RF classification followed by assessment of class membership probabilities	• Class membership probabilities may show significantly elevated values for units that were not predicted as winning class in final classification
Identifying where a classification operating spatially beyond the scope of training data	RF Classification + Assessment of H	• Assess H for suspiciously low values associated with the prediction of a boundary proximal class

References

Cracknell, M. and Reading, A. (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): 113 – 126.

Cracknell, M., Reading, A. and McNeill, A. (2014). Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests^{TM} and Self-Organising Maps, *Australian Journal of Earth Sciences* **61**: 287–304.

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. and Zupan, B. (2013). Orange: Data mining toolbox in python, *Journal of Machine Learning Research* **14**: 2349–2353.

Grunsky, E. and Kjarsgaard, B. (2015). Recognizing and validating structural processes in geochemical data, *Compositional Data Analysis: Springer Proceedings in Mathematics and Statistics* p. 85–116.

Hood, S. (2019). *Linking protolith rocks to altered equivalents by combining unsupervised and supervised machine learning*, Ph.D. Thesis, University of Tasmania.

Hood, S., Cracknell, M. and Gazley, F. (2018). Linking protolith rocks to altered equivalents by combining unsupervised and supervised machine learning, **186**.

Kuhn, S., Cracknell, M. and Reading, A. (2018). Multivariate spatial analysis of lake sediment geochemical data; Melville Peninsula, Nunavut, Canada, *Geophysics* **83**: B183–B193.

Mueller, U. A. and Grunsky, E. C. (2016). Lithological mapping using Random Forests applied to geophysical and remote sensing data: a demonstration study from the Eastern Goldfields of Australia, *Geophysics* **83**: B183–B193.

Pawlowski-Glahn, V. and Buccianti, A. (2012). *Compositional Data Analysis, Theory and Applications*, Wiley and Sons.

Shannon, C. (1948). A mathematical theory of communication, *Bell Systems Technical Journal* **27**: 379–423.

Chapter 7

Conclusions

This thesis comprises a series of case studies demonstrating the efficacy of machine learning for lithological mapping across a range of geological settings. Random Forests, and, to a lesser extent, the clustering algorithms Self-Organising Maps and k-means are deployed for lithological map production or refinement at various stages of project maturity. These case studies demonstrate the ability of Random Forests to rank, prioritise and reduce the number of input variables to those essential for an accurate result, giving explorers insight into useful datasets for mapping which may not have been anticipated via traditional workflows. This allows explorers to extract value from the higher dimensionality search space available to machine learning techniques using only as many variables as is necessary to produce the best result, making interpretation and validation of the end products by geoscientists more feasible. Further interpretation or clustering benefits from this identification of non-redundant variables.

Information entropy (H) is the uncertainty metric found to be of utility throughout this thesis. It provides a monotonic expression of uncertainty; a combination of possible inaccuracy and complex result combinations, or, can be normalised in a form that correlates with an increasing probability of inaccurate classification, as demonstrated in this thesis. This gives explorers the option to omit results at a cut off level of their choosing, biased towards completeness or correctness at their choosing. Random Forests, a soft classifier, allows for quantitative prediction evaluation calculations to test the validity of classification. Moreover, it was found that an elevated class membership probability could be of particular value to explorers in identifying indicator lithologies for subsequent field investigation.

Clustering was successfully used to produce a lithological map of a project area. Results are improved through the omission of redundant variables as identified using Random Forests ranking. Results showed two maxima with an optimal (according to numerical scoring of cluster coherence) small cluster number corresponding to tectonic domains and a larger number better corresponding to lithologies. This makes some understanding of the geological setting, and number of lithology types present, desirable in adding the most value through the machine learning procedure.

In the three studies presented in this thesis, Random Forests is used to perform an objective audit of an existing geological map. This was demonstrated to be an effective means of adding value at any stage of project maturity: from early stages with limited reconnaissance stage data (airborne potential field, remote sensing) or with the full benefit of ground based geochemical sampling. This procedure allows a geological map to be tested in a repeatable and objective fashion and

subsequently refined as needed. By taking enough samples for a robust training dataset, but a proportionally smaller number relative to the total map area, enough freedom is allowed for the lithological map to be reclassified.

All workflows and methods presented in this thesis were selected with the non-computer scientist in mind. The selection of machine learning GUI, industry standard software and the use of geoscience terminology, as opposed to mathematical or data science-oriented terminology, lowers the barrier to entry and thus will enable the much wider uptake of machine learning by mineral explorers.

Appendix A

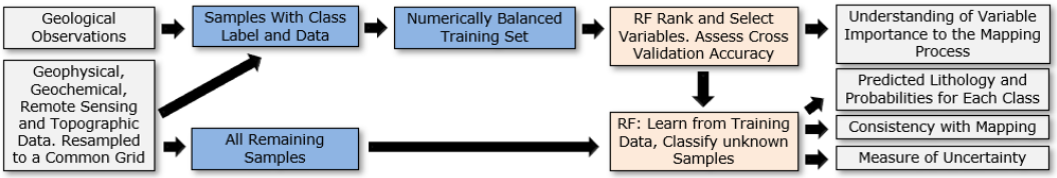
Supplement to Chapter 4

TABLE A.1: All conditionally independent variables considered and ranked in this study, prior to experimental selection of top 15 for RF training (as described in main text Methods).

Geophysical and Remote Sensing Data	
Dataset	**Abbreviation**
Reduced to Pole Total Magnetic Intensity	RTP
RTP - First vertical Derivative	RTP _1vd
Total Magnetic Intensity - Analytic Signal	ASIG
Radiometric - Potassium	K_rad
Radiometric - Thorium	Th_Rad
Radiometric - Uranium	U_Rad
Apparent Resistivity (AEM 33KHz coplanar)	Res33K
Landsat 7: band 4	LSB4
Landsat 7: band 6	LSB6
Landsat 7: band 7	LSB7
DTM (Shuttle Radar Topography mission)	SRTM

GEOCHEMICAL Data		
Ag	Ga	Pb
Al	In	Rb
As	Li	S
Au	Mg	Sb
Ba	Mn	Ti
Ca	Mo	W
Cd	Na	Y
Cu	Nb	Zn
Fe	Ni	Zr

FIGURE A.1: Random Forests Workflow and Modifiable parameters of Random Forest used in classification. Implemented in Orange 3 (Demsar et al. (2013))



Number of Trees: 500

Number of input variables: 15

Measure of split quality used to split at each node: Gini Impurity

Depth limit of individual trees: None

Minimum number of samples required to split node: 2

Number of variables considered at each split: Square root of number of input features

References

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M. and Zupan, B. (2013). Orange: Data mining toolbox in python, *Journal of Machine Learning Research* **14**: 2349–2353.