Ore Geology Reviews 112 (2019) 103015

Contents lists available at ScienceDirect

Ore Geology Reviews

journal homepage: www.elsevier.com/locate/oregeorev

Lithological mapping in the Central African Copper Belt using Random Forests and clustering: Strategies for optimised results

Stephen Kuhn[a,b,*], Matthew J. Cracknell[a,b], Anya M. Reading[b,c]

[a] School of Natural Sciences (Earth Sciences) and CODES Centre of Excellence in Ore Deposits, University of Tasmania, Australia
[b] ARC Industrial Transformation Research Hub for Transforming the Mining Value Chain, University of Tasmania, Australia
[c] School of Natural Sciences (Physics) and CODES Centre of Excellence in Ore Deposits, University of Tasmania, Australia

ARTICLE INFO

Keywords:
Machine learning
Lithological mapping
Uncertainty
Central African Copper Belt

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.

1. Introduction

The Trident project, held by First Quantum Minerals Ltd (FQM) is situated in the North-Western province of Zambia (Fig. 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.

* Corresponding author at: School of Natural Sciences (Earth Sciences) and CODES Centre of Excellence in Ore Deposits, University of Tasmania, Australia.
E-mail address: stephen.kuhn@utas.edu.au (S. Kuhn).

https://doi.org/10.1016/j.oregeorev.2019.103015
Received 3 May 2018; Received in revised form 10 April 2019; Accepted 11 July 2019
Available online 12 July 2019

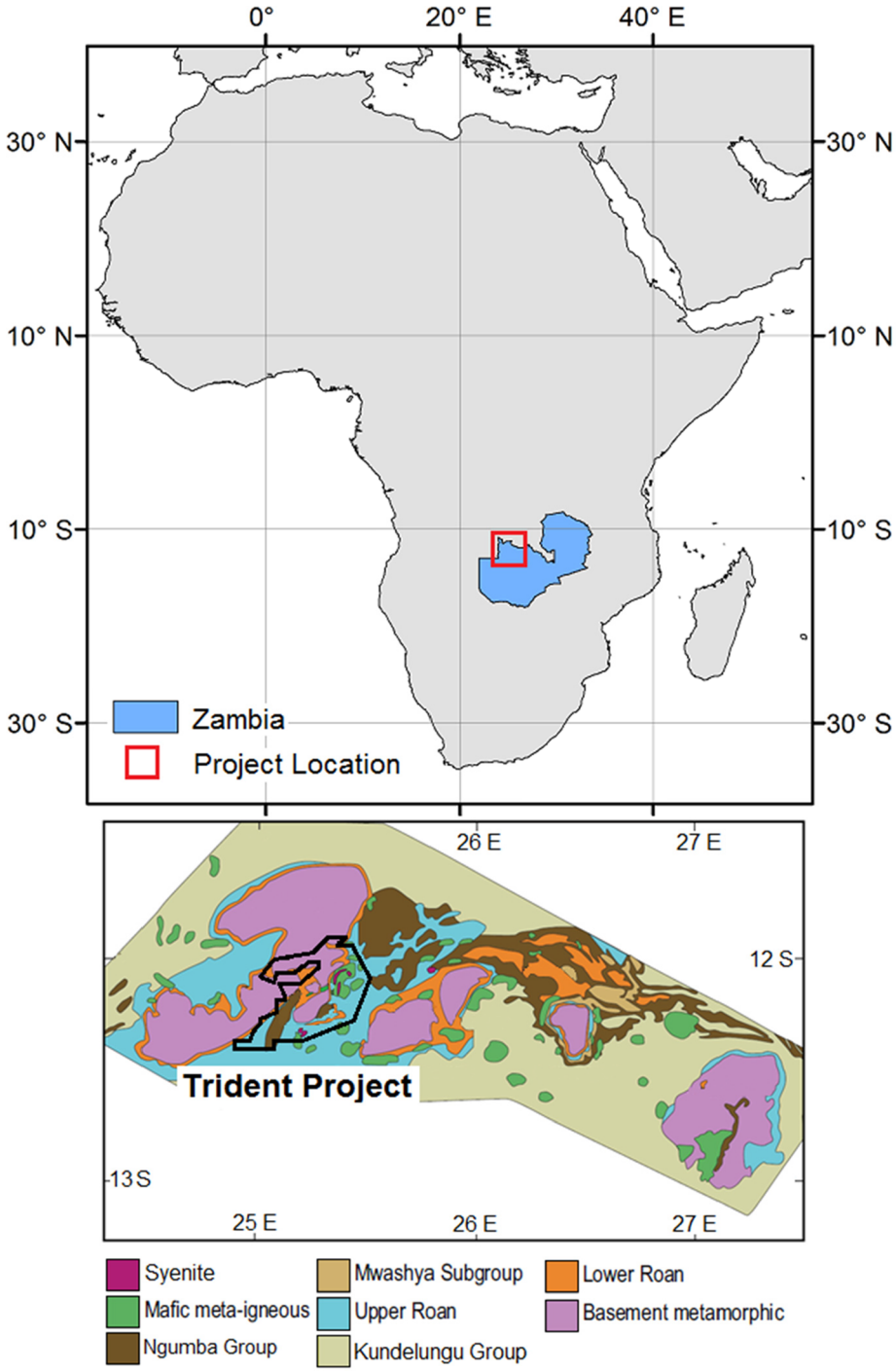

Fig. 1. (Top) Project location relative to the African Continent and the country of Zambia. (Bottom) Schematic summary geology of northern North-Western Zambia (modified from Capistrant et al., 2015) showing the location of the Trident project (red outline). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1.1. Geology

The Trident project area is approximately 75 km × 40 km in size and located in the Domes region of the CACB. The region, as described by Capistrant et al. (2015; Fig. 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 (Fig. 2). This map has undergone several updates that

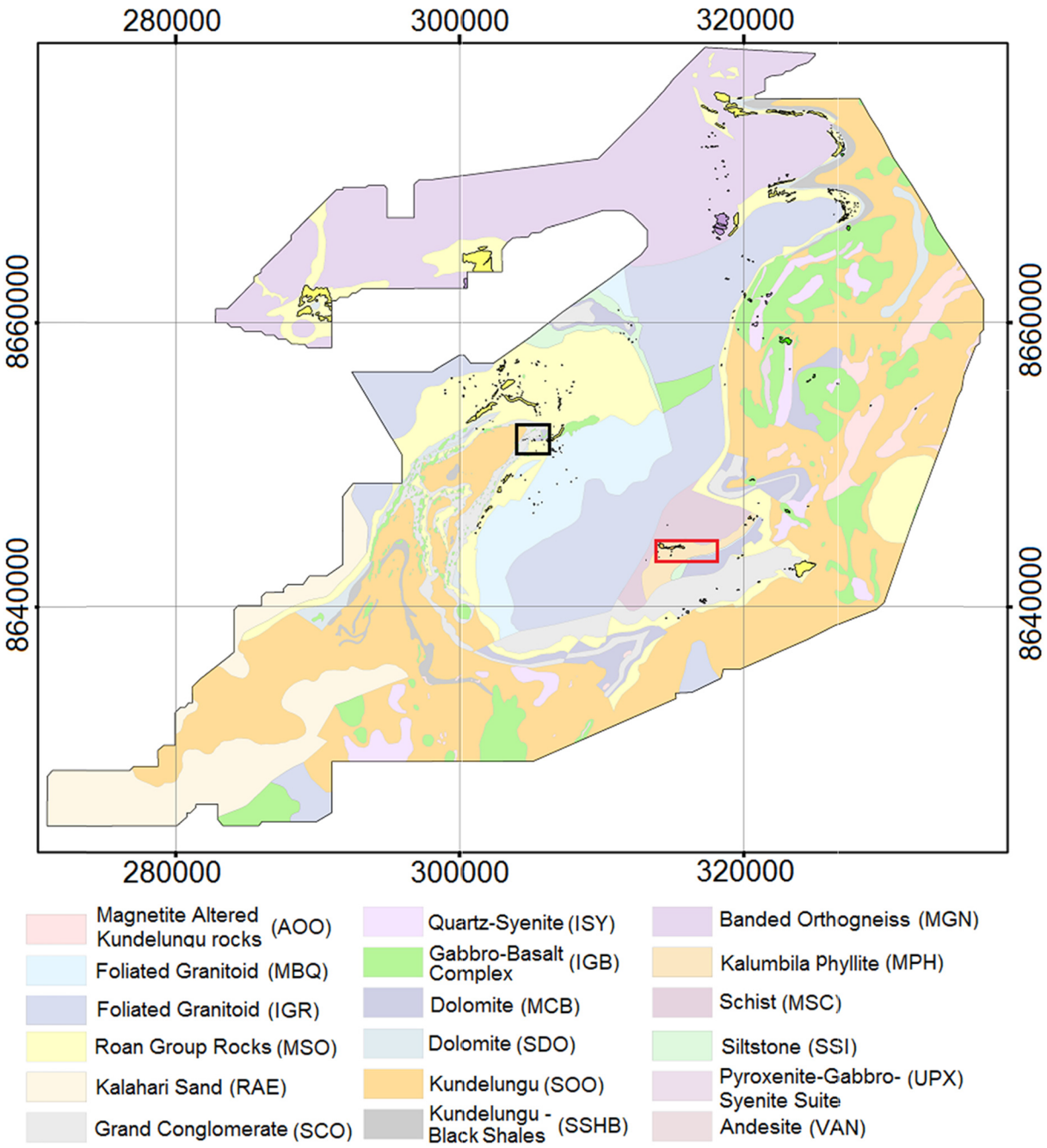

Fig. 2. Initial map of interpreted lithology under cover (pale colours) showing outcrop locations (solid colours). The Enterprise and Sentinel deposits are located within the black and red boxes respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

combine well-defined stratigraphy, as described above, and lithologies with an unconfirmed stratigraphic position. The 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).

1.2. Random Forests

Random Forests™ (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. Liaw and Wiener, 2002; Hastie et al., 2009; Kuhn et al., 2016, 2018) assembles a 'forest' comprising many classification trees (Fig. 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, trees are built on randomly selected subsets of training data (T_a) 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

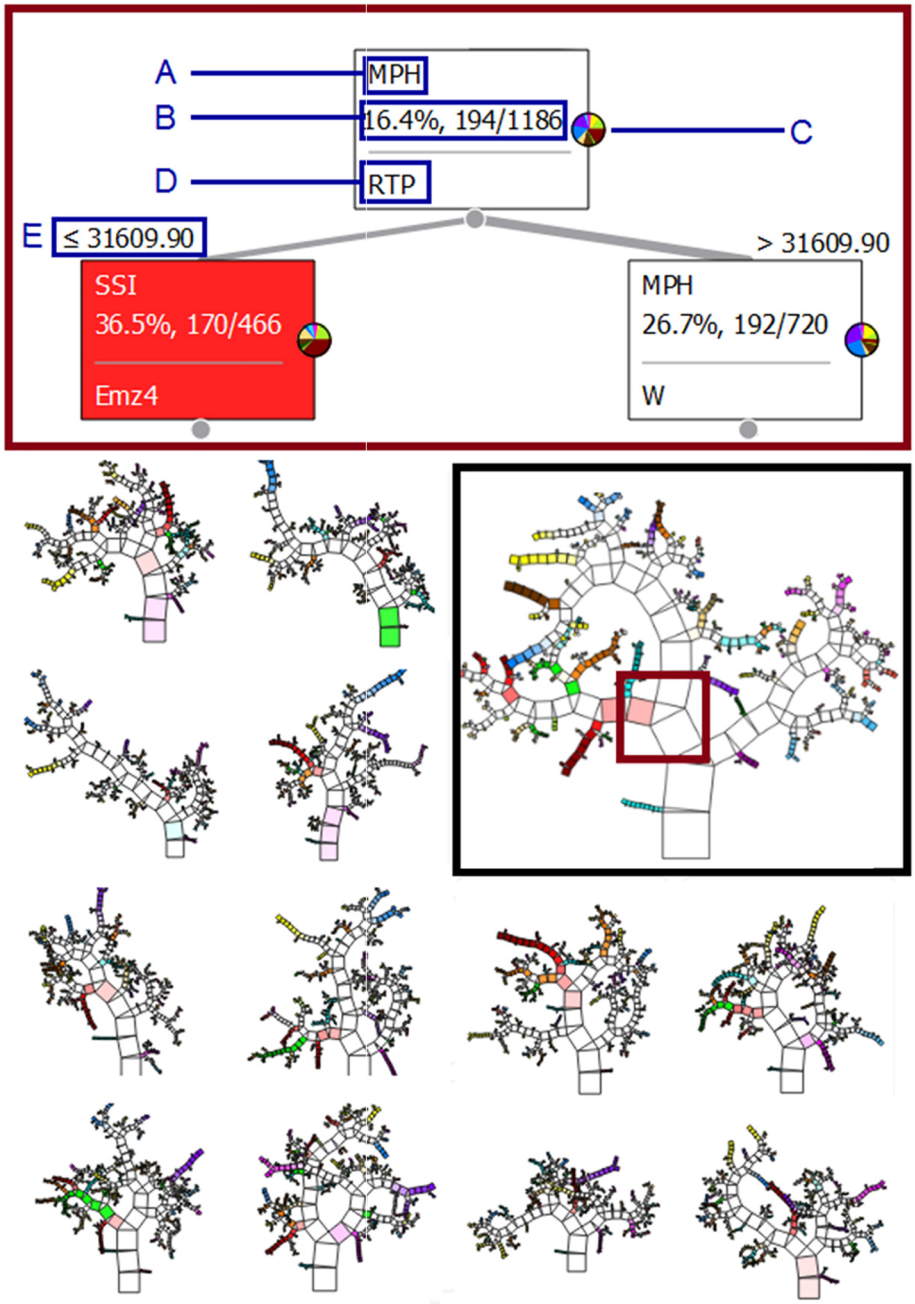

Fig. 3. Schematic example from a RF used in this study highlighting an example of a node split (red box) where A is the nodes dominant class, B is the proportion as percent and count of the node that class occupies, C is the spread of classes also shown as a pie chart, D is the variable used to split 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

improvement in node homogeneity as defined by decrease in Gini index (Breiman et al., 1984), is selected to split that node (Fig. 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.

1.3. 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:

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

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 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 (Cracknell and Reading, 2013; 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).

1.4. Clustering

The k-means algorithm (Lloyd, 1957) 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, 2006) 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.

The Self-Organising Maps (SOM) algorithm, 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; Bierlein 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.

1.5. 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.

2. Data and methods

2.1. Data compilation and pre-processing

Data used in this study were provided by FQM. These comprise both geophysical and soil geochemical data (Table 1; Fig. 4). Additional geophysical datasets were derived from those provided and the Shuttle Radar Topography Mission (SRTM; National Aeronautics and Space Administration, 2003) digital terrain model (DTM) was added. 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 × 300 m) were gridded to a

Table 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	

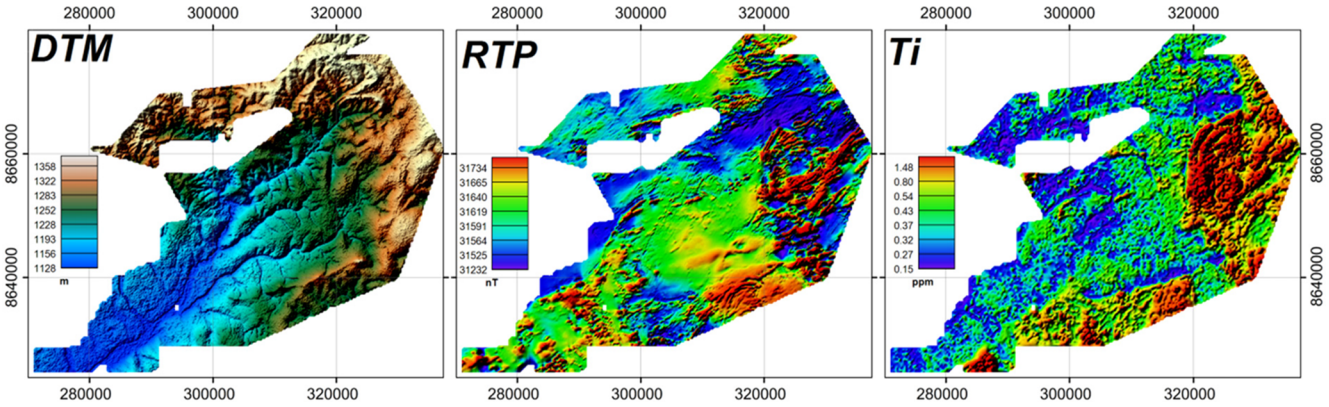

Fig. 4. Examples of 3 variables used in this study: DTM, RTP magnetics and Ti. These variables were deemed useful in case studies C1, C2 and C3.

100 m cell size. All data were resampled to a regular grid of 100 m × 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.

2.2. Variable prioritisation and reduction

2.2.1. 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 1).

2.2.2. 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; Demsar et al., 2013). 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.

2.3. Sampling

2.3.1. Case study 1

Case study one (C1, early exploration stage) used mapped outcrop locations as training data (Fig. 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 (Fig. 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.

2.3.2. 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 2). A sample size of 100 per class was deemed to provide this balance of adequate sample size while introducing an acceptable amount of synthetic samples (Fig. 5B). Larger sample sizes retained more real data but introduced an unacceptably high proportion of synthetic samples across all represented classes.

2.3.3. 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 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 (Fig. 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.

2.3.4. 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

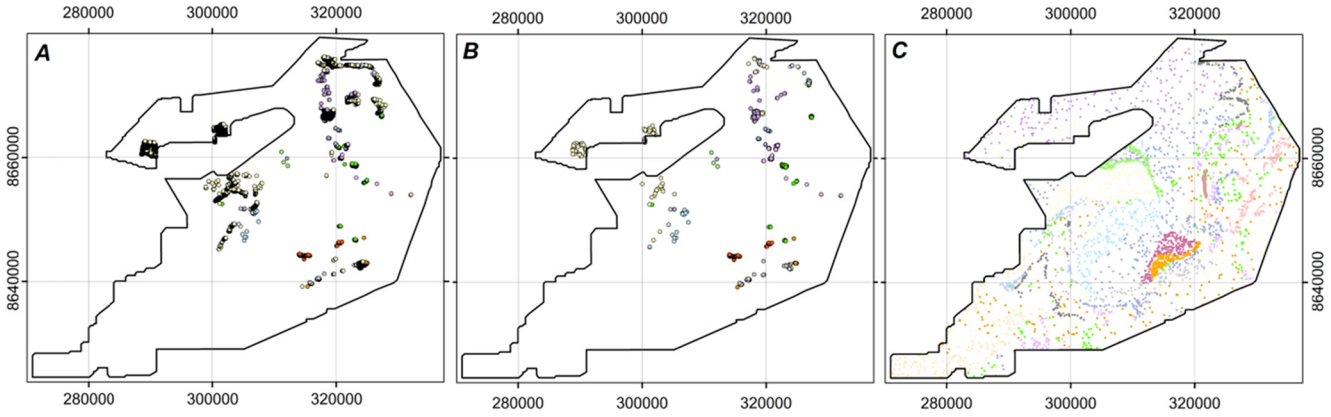

Fig. 5. Training data locations for (A) case study C1, (B) C2 and (C) C3. Note the diameter of each sample in (A) and (B) has been increased by a factor of 7 and in (C) by a factor of 3, for legibility. See Fig. 2 for lithology colour key.

Table 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

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 (Fig. 6). As such 300 iterations were used for all k-means models. SOM parameters including map size and dimensions, were investigated and a 45×45 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.

3. Results

3.1. 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 3; Fig. 7). Ranking of training data from C2 defined 10 relevant variables (Table 3; Fig. 7), producing a peak cross validation accuracy of 88.8%. Ranking of datasets using training data from C3 identified 12 relevant variables (Table 3; Fig. 7), producing a peak cross validation accuracy of 80.5%.

3.2. 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 (Fig. 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 (Fig. 8B).

3.3. C2 classification results

As with C1, consistency with respect to the geological interpretation map was poor. In this case however, while lithology labels remain inaccurate, the balanced training sample produced results that more closely conform to the geometry of major boundaries in the project area (Fig. 9A). Higher values of H in this case (Fig. 9C) show a spatial relationship with lithological boundaries and areas of geological complexity in regions approximating the spatial range of the training data. 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 (Fig. 9A; DTM in Fig. 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

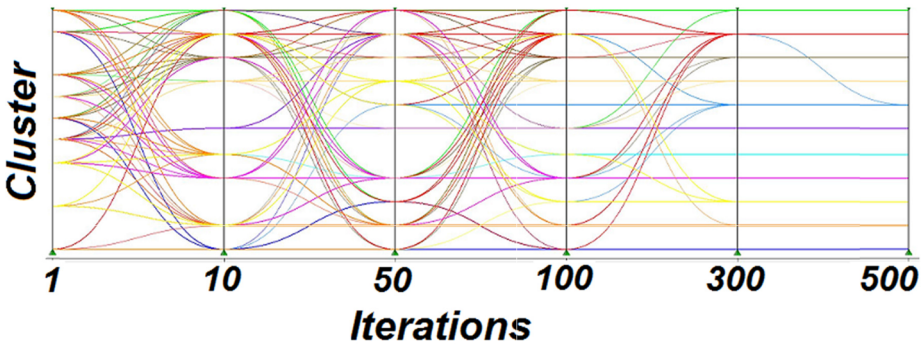

Fig. 6. k-means convergence vs iterations performed. Lines represent the assignment and subsequent reassignment/refinement of samples as the number of iterations is increased. Lines are smoothed between experiments and reflect the reassignment path (and not the assignment of samples at iteration increments between those displayed).

Table 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 cutoff 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

in a lithology prediction that saw better recovery of interpreted boundary geometries (Fig. 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 (Fig. 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.

3.4. 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 (Fig. 10A) with 67.2% consistency with respect to that map (Fig. 10B). The confusion matrix associated with this classification (Table 4) shows that 10 of the 17 lithological classes achieved a recall in excess of 75% and a further three classes above 65%. 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 4). This can be seen clearly in the expression of class membership probabilities, examples of which are shown in Fig. 11. The MGN lithology class (Fig. 11E, predicted with a recall of 96%, exhibits very concise and distinct regions where this class was probable. Conversely, the lithology class SOO (Fig. 11D) was not predicted by a large majority, as shown by relatively low class membership probabilities across its spatial range. H (Fig. 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} (Fig. 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 (Fig. 13).

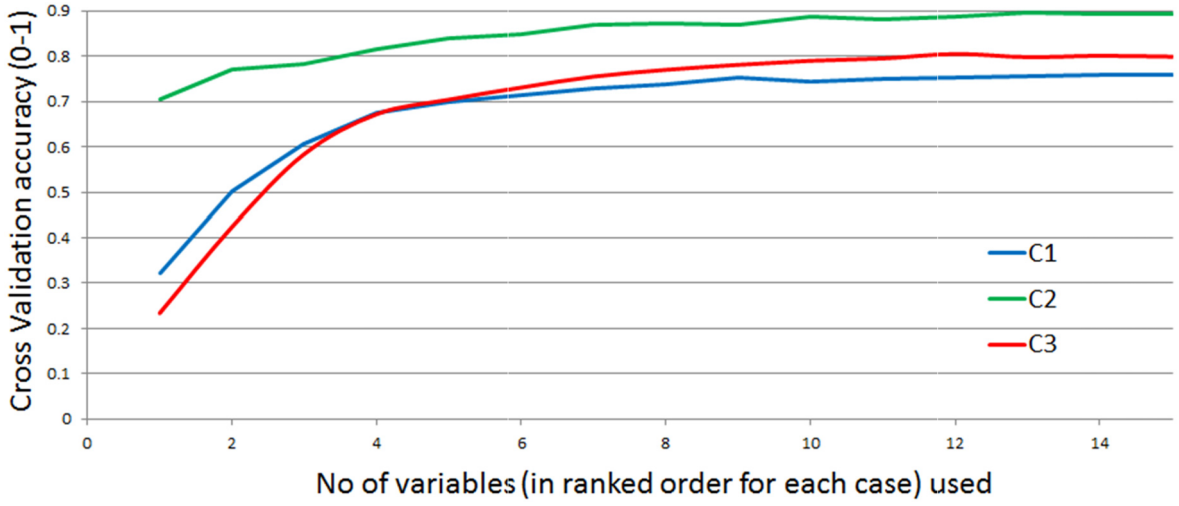

Fig. 7. Cross validation accuracy with addition of successively lower ranked variables for each RF case study. (C1) sampling from outcrop, (C2) class size balanced sampling from outcrop and (C3) sampling from a geological map. Note the accuracy using balanced outcrop-based sampling (C2) is strongly influenced 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.

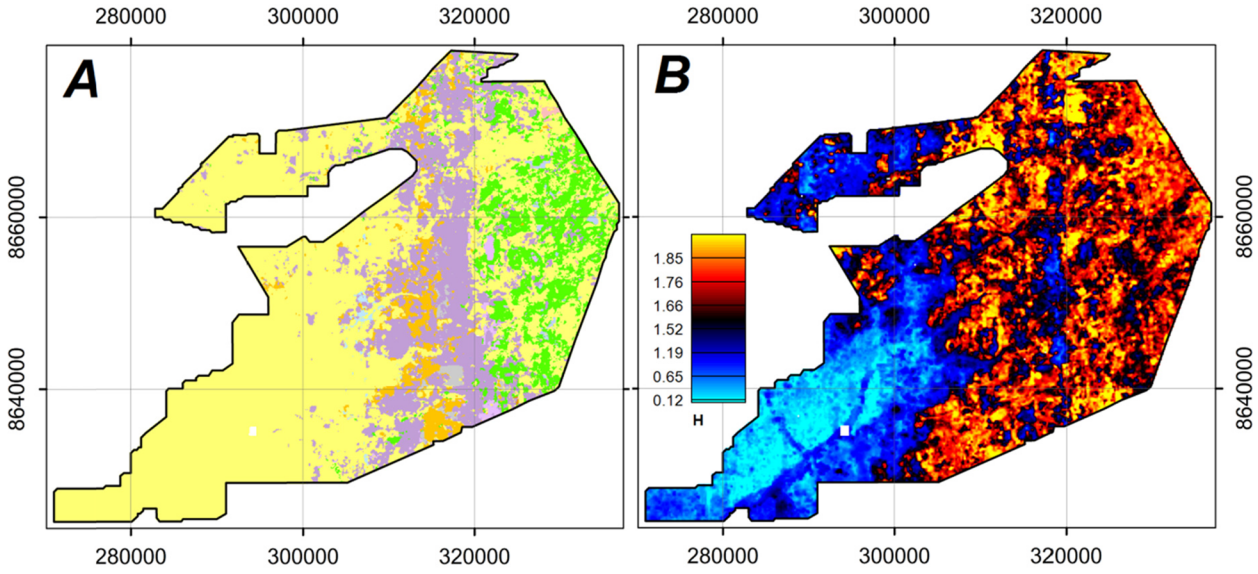

Fig. 8. (A) Classification output using C1 training data. See Fig. 2 for lithology colour key. (B) H associated with C1 classification output. Note that in addition to poor accuracy with respect to interpreted lithology on a pixel by pixel basis, interpreted geometry and structure are absent, in favour 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.

3.5. Clustering results

Using k-means and SOM-CL produced a series of outputs of 2–20 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 (Fig. 14).

4. Discussion

4.1. 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 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, 1993). 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 T_a, 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 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 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.

4.2. RF 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 (Fig. 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

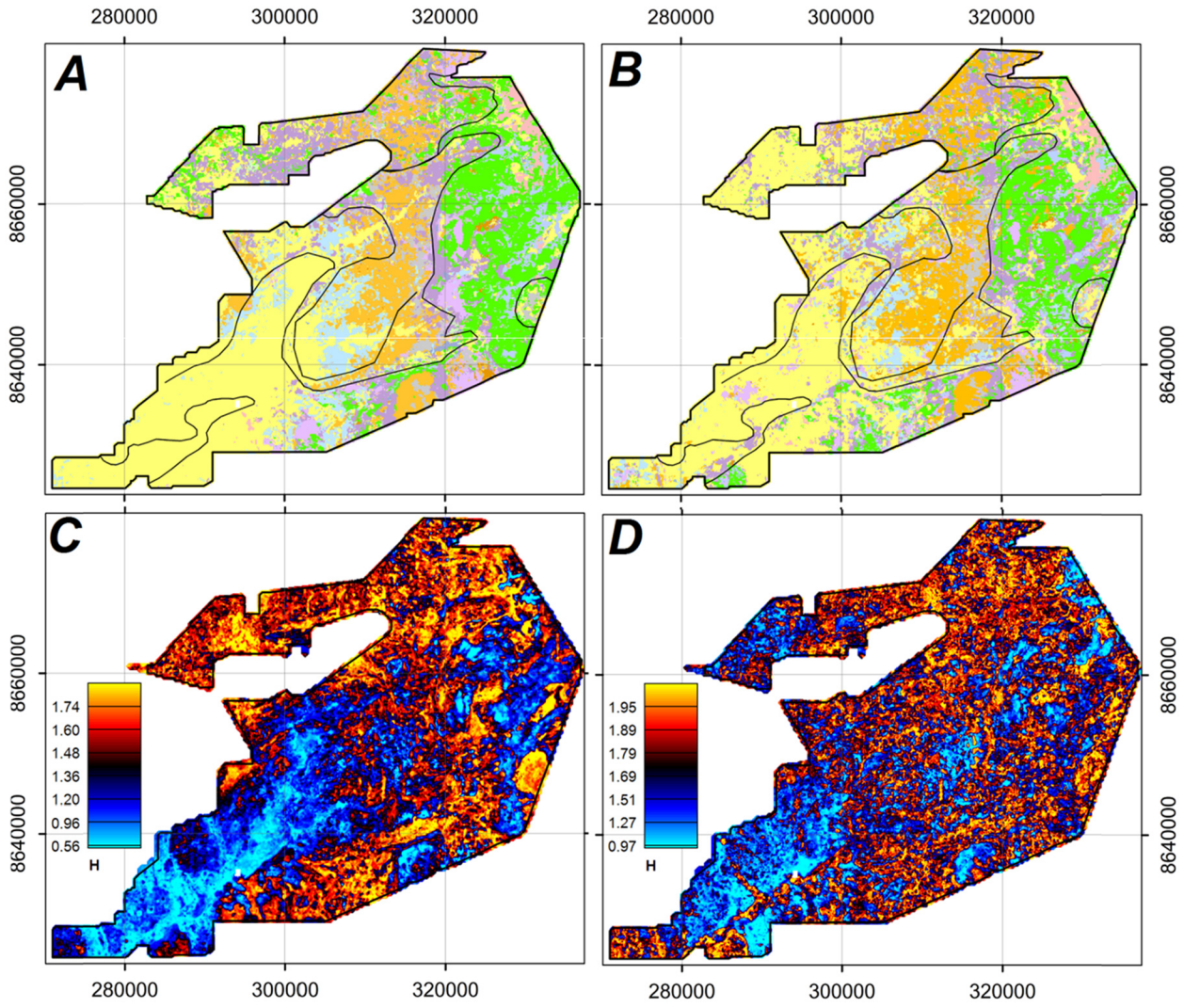

Fig. 9. (A) Classified lithology map refined using C2 training data. See Fig. 2 for lithology colour key. (B) Classified lithology map using C2 training data adjusted to omit the DTM. (C) H associated with (A). (D) H associated with (B). Note that while lithology prediction accuracy is poor on a per pixel basis, major geometries/boundaries are present.

training data, were predicted in a more geological reasonable manner (Fig. 9A and B). 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 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, Fig. 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 T_a (outcrop), prediction of the class label of the nearest training data was common. This occurred most notably, in the southwest (Figs. 8A, 9A and 9B). These predictions were associated with anomalously low H (Figs. 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, 2016; Cracknell and Reading, 2014; Cracknell, Reading and McNeill, 2014) H and H_{norm} associated with this bulk, incorrect or 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 T_a used. This indicates a spatial transition into an unsampled geological domain but could also occur when presented with rock types not included in the T_a, regardless of spatial range.

4.3. 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

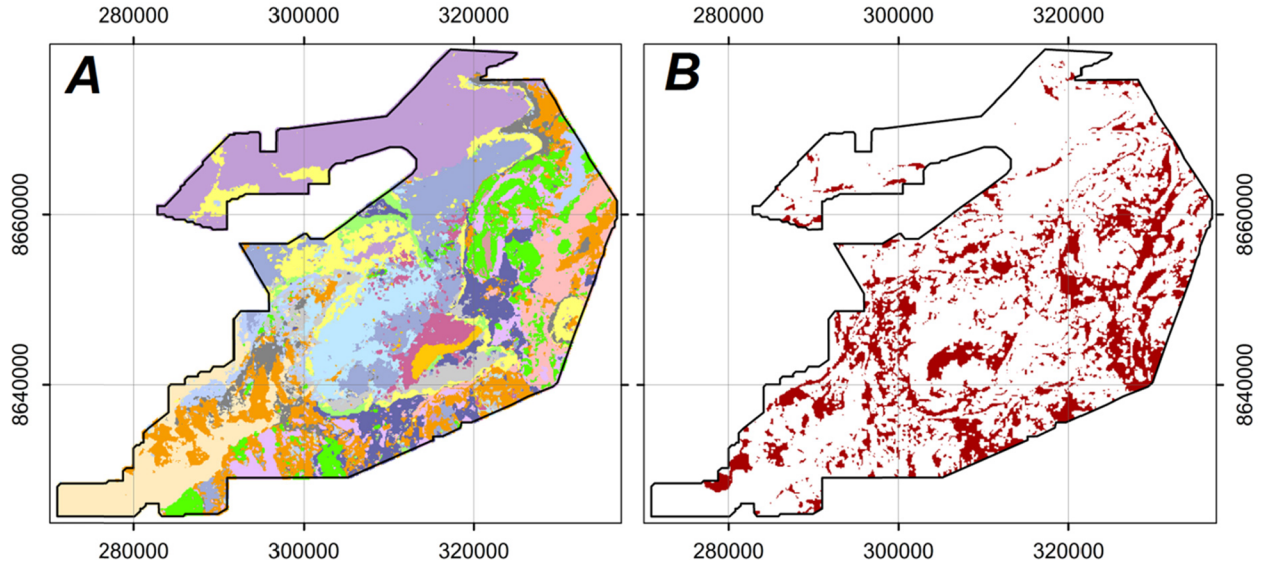

Fig. 10. (A) Classified lithology map refined using training data C3. See Fig. 2 for lithology colour key. (B) Comparison with the initial map of interpreted geology (Fig. 2) as consistent (white) and inconsistent (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

moderate, at 67.5% (Fig. 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_{norn}.

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 (Fig. 11) can be used to better understand the confidence in prediction

Table 4
C3 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 Fig. 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
IGB	10	62	2	6	1	3	0	0	0	0	1	3	2	4	2	0	4
IGR	0	1	68	0	13	1	0	0	7	3	0	2	1	3	0	0	0
ISY	3	11	0	79	0	3	0	0	0	0	0	0	0	3	1	0	0
MBQ	0	0	7	0	81	0	0	0	5	2	0	3	0	0	0	1	0
MCB	2	2	0	2	0	79	0	1	2	1	0	5	0	2	0	3	0
MGN	0	0	1	0	0	0	96	0	0	3	0	0	0	0	0	0	0
MPH	0	0	0	0	0	2	0	91	3	0	0	1	0	0	0	3	0
MSC	0	0	1	0	2	0	0	11	85	0	0	1	0	0	0	0	0
MSO	1	1	5	0	6	2	10	0	2	54	1	3	5	2	3	5	0
RAE	0	1	0	0	0	0	0	0	0	0	95	0	2	1	0	0	0
SCO	0	1	2	1	5	13	0	0	3	7	3	52	4	4	2	2	0
SDO	0	2	0	0	2	0	0	0	0	7	3	3	67	2	13	0	0
SOO	10	7	1	9	2	11	1	0	0	1	7	3	4	38	5	0	0
SSHB	1	3	0	1	0	0	0	0	0	1	12	0	5	2	74	0	0
SSI	0	0	1	0	0	2	0	1	1	2	0	0	0	0	0	92	0
UPX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
Count	9121	12514	17459	8300	11562	10355	28095	1352	4513	14030	14795	7076	5699	21049	5379	2589	594

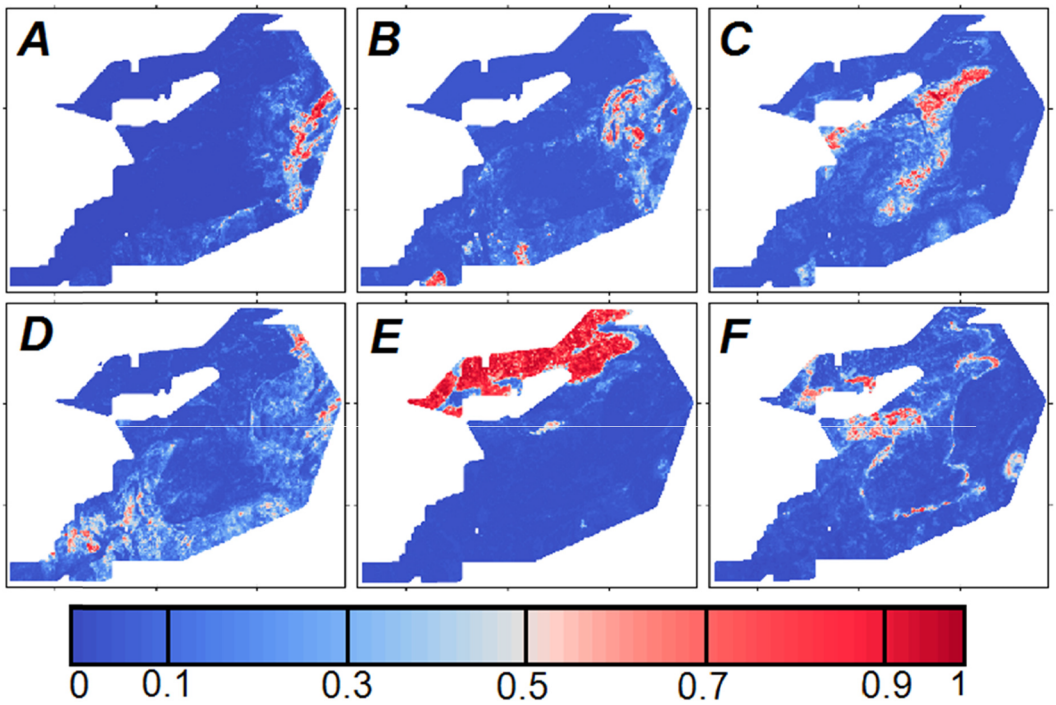

Fig. 11. Examples of case study C3 class membership probabilities. (A) AOO, (B) IGB, (C) IGR, (D) SOO, (E) MGN and (F) MSO. Rock codes are given in Fig. 2.

of lithology on a per-unit basis. Quantified uncertainty, in the form of H (Fig. 12A) and H_{norm} (Fig. 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 (Fig. 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 (Fig. 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 Fig. 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

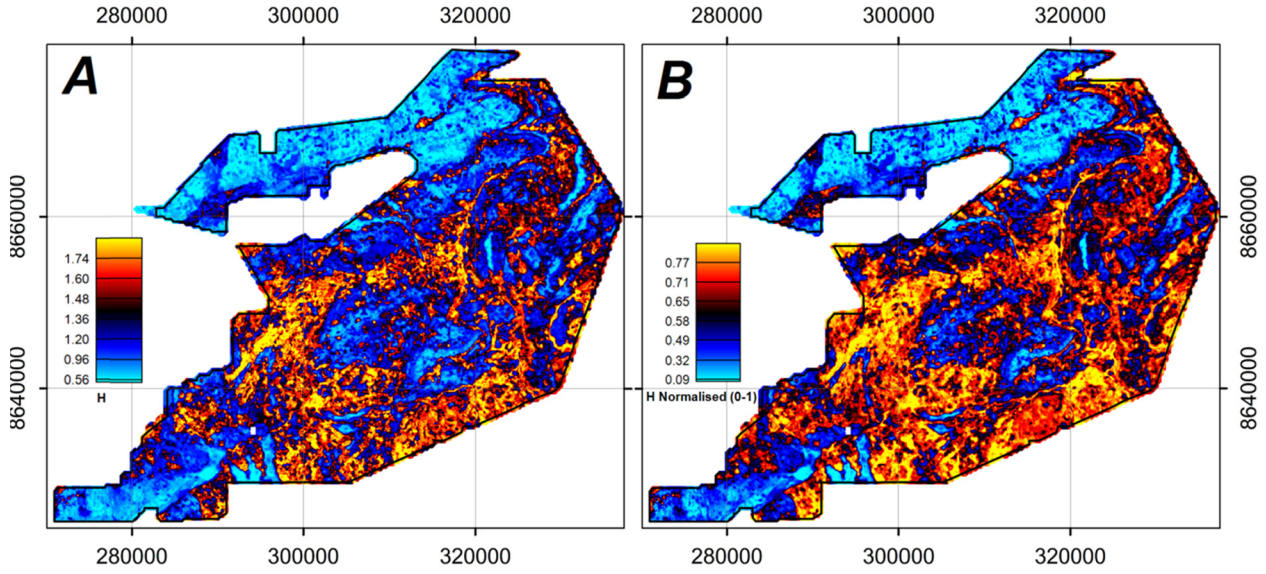

Fig. 12. Case study C3: (A) H and (B) H normalised to 0–1. H is an indication of the disorder at a given point and rises as complexity, i.e. the number of possible classes, increases. When normalised, H provides an indication of how closely a given pixel reaches its maximum possible state of disorder. As such, pixels can be compared and can be a better proxy for prediction accuracy.

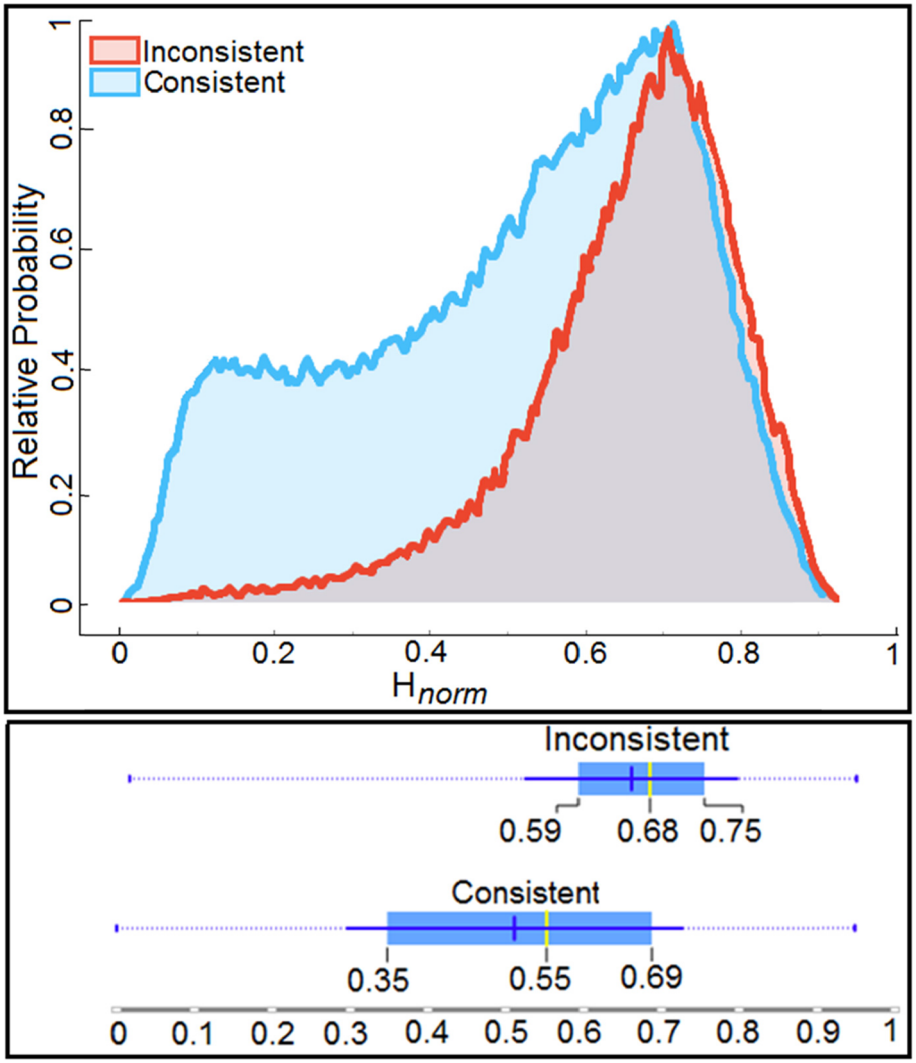

Fig. 13. The distribution of H for C3 partitioned into two groups: samples classified consistently, or inconsistently, relative to the initial interpreted lithology map (Fig. 2). (Top) The relative probability of a consistent or inconsistent classification for any given H_{norm}. (Bottom) Box plot showing the distribution of H_{norm} for consistent and inconsistently classified sample populations. Note that at above a H_{norm} 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 H_{norm} of 0.5, a consistently classified sample is considerably more probable.

higher than was the case for C1 and C2, the absolute range of H is not comparable across the 3 studies.

4.4. 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 (Fig. 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 (Fig. 14: k-means cluster 11 and SOM-CL cluster 10).

K-means is relatively easy to implement 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. 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

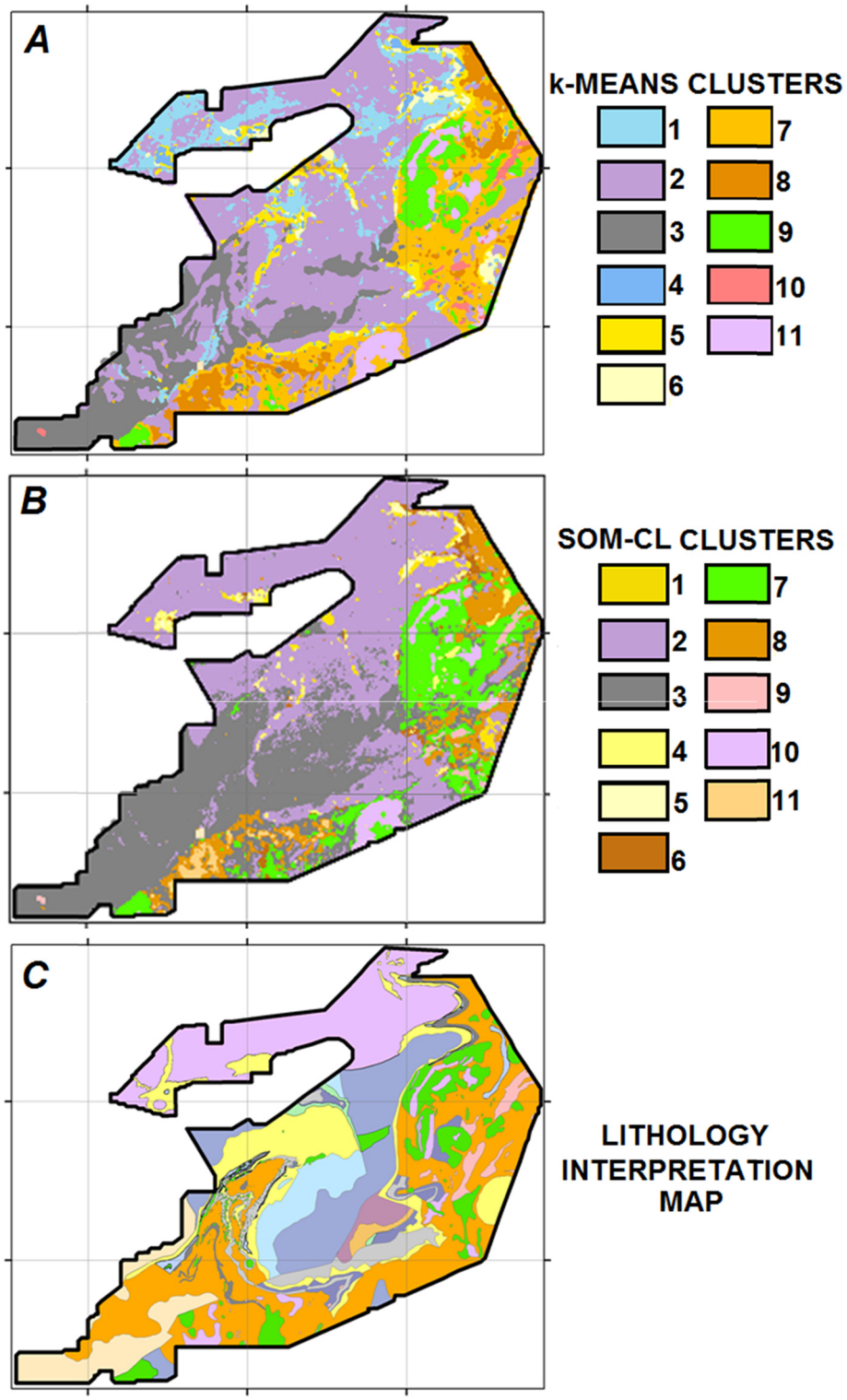

Fig. 14. Comparison of lithology maps. (A) Generated by clustering using k-means and (B) generated by clustering using SOM-CL. (C) The initial interpreted lithology map (Fig. 2) is replotted at the same scale to facilitate a visual comparison (C). Clusters are coloured for the best comparison for that clustering output with initial mapped lithology.

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 T_a 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 T_a. 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.

Acknowledgments

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.oregeorev.2019.103015.

References

Arthur, D., Vassilvitskii, S., 2006. k-means + +: the advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms.

Bank of Zambia 2016, Bank of Zambia Annual Report, 2016, accessed October, 2017, http://www.boz.zm/annual-reports.htm.

Beck, F., Burch, M., Munz, T., Di Silvestro, L., Weiskopf, D., 2014. Generalized Pythagoras trees for visualizing hierarchies. 9th International Conference on Information Visualisation Theory and Applications.

Bierlein, F., Fraser, S., Brown, W., Lees, T., 2008. Advanced methodologies for the analysis of databases of mineraldeposits and major faults. Aust. J. Earth Sci. 55 (1), 79–99.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

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

Capistrant, P.L., Hitzman, M.W., Kelly, N.M., Kuiper, Y., Wood, D., Williams, G., Zimba, M., Jack, D., Stein, H., 2015. Geology of the enterprise hydrothermal nickel deposit, North-Western Province, Zambia. Econ. Geol. 110 (1), 9–38.

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

Cracknell, M., Reading, A., De Caritat, P., 2015. Multiple influences on regolith characteristics from continental-scale geophysical and mineralogical remote sensing data using Self-Organizing Maps. Remote Sens. Environ. 165.

Cracknell, M.J., Reading, A.M., McNeill, A.W., 2014. Mapping geology and volcanic-hosted massive sulphide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests™ and self-organising maps. Aust. J. Earth Sci. 61, 287–304.

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

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2), 224.

Defays, D., 1977. An efficient algorithm for a complete link method. Comput. J. 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., Zupan, B., 2013. Orange: data mining toolbox in python. J. Mach. Learn. Res. 14, 2349–2353.

Fraser, S.J., Dickson, B.L., 2008. A new method for data integration and integrated data interpretation: self-organising maps. 5th Decennial International Conference on Mineral Exploration.

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

Harris, J.R., Grunsky, E.C., 2015. Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Comput. Geosci. 80, 9–25.

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

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 56–69.

Kohonen, T., 2001. Self-Organizing Maps: Springer Series in Information Sciences. Springer-Verlag, Berlin.

Kuhn, S., Cracknell, M.J., Reading, A.M., 2018. Lithological mapping using Random Forests applied to geophysical and remote sensing data: a demonstration study from the Eastern Goldfields of Australia. Geophysics published online ahead of print (accessed 20/04/2018).

Kuhn, S., Cracknell, M.J., Reading, A.M., 2016. Lithological mapping via random forests: information entropy as a proxy for inaccuracy. ASEG-PESA-AIG: 25th International Geophysical Conference and Exhibition. . https://doi.org/10.1190/geo2017-0590.1.

Liaw, A., Wiener, M., 2002. Classification and regression by random Forest. R News 2, 18–22.

Lloyd, S., 1957. Least squares quantization in PCM: Technical Note, Bell Laboratories. Publ. IEEE Trans. Inform. Theory 28 (2), 129.

MacLean, W.H., Barrett, T.J., 1993. Lithogeochemical techniques using immobile elements. J. Geochem. Explorat. 48 (2), 109–133.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley. pp. 281–297.

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

Pearce, J.A., Norry, M.J., 1979. Petrogenetic implications of Ti, Zr, Y, and Nb variations in volcanic rocks. Contrib. Mineral. Petrol. 69 (1), 33–47.

Penn, B.S., 2005. Using self-organizing maps to visualize high-dimensional data. Comput. Geosci. 31 (5), 531–544.

Rodriguez-Galiano, V.F., Chica-Olmo, M., 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: Int. J. Geogr. Inform. Sci. 28 (7), 1336–1354.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Selley, D., Broughton, D., Scott, R.J., Hitzman, M., Bull, S.W., Large, R.R., McGoldrick, P.J., Croaker, M., Pollington, N., 2005. A new look at the geology of the Zambian Copperbelt. In: **One Hundredth Anniversary Volume**, paper 28, pp. 965–1000.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.

Ultsch, A., Vetter, C., 1994. Self-Organising Feature Maps versus Statistical Clustering A Benchmark. Department of Mathematics and Computer Science, University of Marburg.

United States Geological Survey, 2003. National Aeronautics and Space Administration Shuttle Radar Topography Mission, 3 Arc Second, Scene S32E121.SRTMGL3; USGS,

Sioux Falls.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc.
Vapnik, V.N., 1998. Statistical Learning Theory. John Wiley Sons Inc.
Waske, B., Benediktsson, J.A., Árnason, K., Sveinsson, J.R., 2009. Mapping of hyperspectral aviris data using machine-learning algorithms. Can. J. Remote Sensing 35, 106–116.
Wehrens, R., Buydens, L.M.C., 2007. Self- and super-organising maps in R: the kohonen package. J. Stat. Softw. 21, 1–19.
Wellmann, J.F., Regenauer-Lieb, K., 2012. Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models. Tectonophysics 526–529, 207–216.