THE UTILITY OF MACHINE LEARNING IN IDENTIFICATION OF KEY GEOPHYSICAL AND GEOCHEMICAL DATASETS: A CASE STUDY IN LITHOLOGICAL MAPPING IN THE CENTRAL AFRICAN COPPER BELT

Stephen Kuhn*
Sch. of Physical Sci. (Earth Sciences)
CODES Centre of Excellence and
ARC ITR Hub for Transforming
the Mining Value Chain
University of Tasmania
stephen.kuhn@utas.edu.au

Matthew J. Cracknell
Sch. of Physical Sci. (Earth Sciences)
CODES Centre of Excellence and
ARC ITR Hub for Transforming
the Mining Value Chain
University of Tasmania
m.j.cracknell@utas.edu.au

Anya M. Reading
Sch. of Physical Sci. (Earth Sciences)
CODES Centre of Excellence and
ARC ITR Hub for Transforming
the Mining Value Chain
University of Tasmania
anya.reading@utas.edu.au

**Presenting Author*

SUMMARY

Random Forests, a supervised machine learning algorithm, provides a robust, data driven means of predicting lithology from geophysical, geochemical and remote sensing data. As an essential part of input selection, datasets are ranked in order of importance to the classification outcome. Those ranked most important provide, on average, the most decisive split between lithological classes. These rankings provide explorers with an additional line of reasoning to complement conventional, geophysical and geochemical interpretation workflows. The approach shows potential to aid in identifying important criteria for distinguishing geological map units during early stage exploration. This can assist in directing subsequent expenditure towards the acquisition and further development of datasets which will be the most productive for mapping.

In this case study, we use Random Forests to classify the lithology of a project in the Central African Copper-Belt, Zambia. The project area boasts extensive magnetic, radiometric, electromagnetic and multi-element geochemical coverage but only sparse geological observations. Under various training data paradigms, Random Forests produced a series of varying but closely related lithological maps. In this study, training data were restricted to outcrop, simulating the data available at the early stages of the project. Variable ranking highlighted those datasets which were of greatest importance to the result. Both geophysical and geochemical datasets were well represented in the highest ranking variables, reinforcing the importance of access to both data types. Further analysis showed that in many cases, the importance of high ranking datasets had a plausible geological explanation, often consistent with conventional interpretation. In other cases the method provides new insights, identifying datasets which may not have been considered from the outset of a new project.

Key words: Random Forests, Machine Learning, Dataset Ranking, Lithological Mapping

INTRODUCTION

Machine learning is seeing increasing uptake in the search for solutions to geoscientific problems. Random Forests (RF) (Breiman, 2001) has proven a strong choice for geological classification problems (Cracknell & Reading, 2014, Kuhn, Cracknell & Reading, 2016; Cracknell, Reading & McNeill, 2014).

As a supervised classification algorithm, RF employs a set of user defined training data to build a set of rules which in turn are used to classify unknown samples. RF is an extension of the classification tree method which mitigates the effects of over fitting by generating a multitude, or forest, of classification trees (Figure 1), each built on a subset of training data.

Randomness is introduced at two stages. Firstly, bootstrap aggregation, also known as bagging, is used to select the subset of training data available to each tree. Secondly, a subset of input variables is selected at random for each node, with the variable from within that subset which provides the best split, used to split that node. Nodes are split to produce the maximum improvement in homogeneity in the child nodes, relative to the parent node until nodes become homogenous to within a defined tolerance.

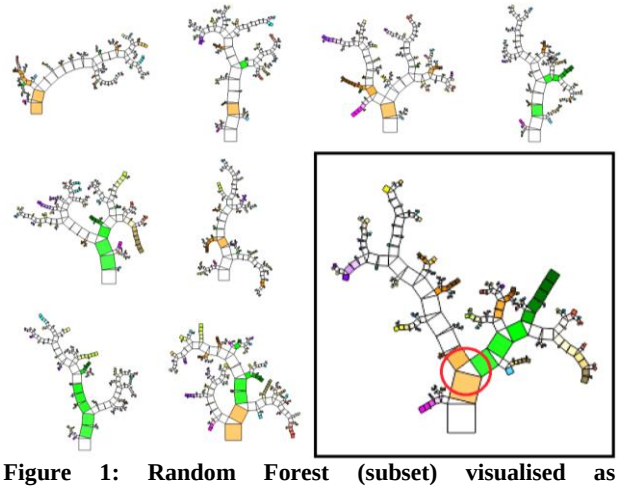

Figure 1: Random Forest (subset) visualised as Pythagorean trees (Beck et. al., 2014). The side length of each square (node) is proportional to the sample size at that node. Side lengths of triangles between nodes describe the partition of samples from parent to child nodes. The enlarged tree shows an example high order split based on Ta (red circle), prior to further class separation. Colours represent the majority class (if present) at each node.

Each tree within the forest casts a vote on the final classification produced by the forest. RF has been shown to achieve comparable accuracy to optimal use of other machine learning algorithms for lithological classification applications (Cracknell & Reading, 2014).

When applied to lithological mapping, a balance must be negotiated between the information gained through the introduction of higher dimensionality and the ability to produce output that is understandable to a human user. Previous studies (Cracknell, Reading & McNeill, 2014; Kuhn, Cracknell & Reading 2016) have shown a tendency for a point of diminishing returns whereby the addition of more input variables no longer produces an improvement in classification results. In order to perform this dimensionality reduction, input variables must be ranked in order of their importance to the RF classification. This process is described in detail by Breiman (2001) and involves a permutation of each input variable through the forest. Those variables which, when permuted produce the largest variations in classification error. While the primary purpose of such ranking is input selection for a RF classification exercise, these rankings are also valuable in isolation, providing an objective starting model for any multi-dataset interpretation.

In this case study, comprising the Trident project area (Figure 2), we used RF to rank a combination of geophysical, geochemical and remote sensing data; provided by First Quantum Minerals Ltd. The Trident project is located within the Domes region of the Central African Copper Belt, in northwest Zambia (Capistrant et. al., 2015). In this example, we limited training data to regions of mapped outcrop (Figure 2) in a simulation of training constraints available to an exploration team prior to the development of additional interpretation.

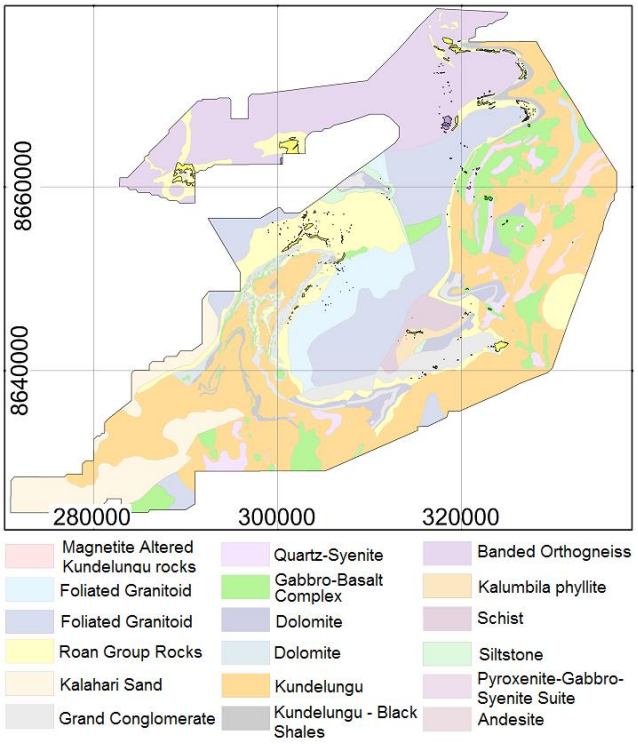

Figure 2. Geology of the Trident project area. Interpreted geology map with observed outcrop, defining the training data used in this study, shown opaque. Map is projected in WGS84 UTM zone 35s.

METHOD AND RESULTS

Data

Data were compiled over the extent of the project area and combined into a single matrix in the form of x, y, d_1, d_2, …, d_n (where x and y are co-ordinates and d are data channels). Geophysical data comprised reduced to pole (RTP) airborne magnetics, radiometrics, multiple airborne EM channels and elevation (DTM). Additional products were derived from these datasets. Soil geochemistry (ICP-MS) was collected at 300 x 300 m spacing. All data were resampled to a 100 x 100 m grid size. As a preliminary dimensionality reduction stage, highly correlated variables ($r > 0.8$) were omitted. With correlated variables removed, a total of 44 data channels were presented to the RF classifier, the top 10 of which are shown in Table 1. Sample size was balanced such that an equal number of samples were available from each lithological class.

Results of Random Forests Ranking

A RF comprising 500 trees (example subset in Figure 1) was used for this exercise with no limitations on depth extent or pruning. Variables were ranked by RF and passed to an additional step of 10 fold cross validation where the classifier performance was tested as variables were successively added. In this case, the top ranked variable would be used in isolation, followed by the top 2 ranked, top 3 and so on. In this study, a point of diminishing returns was reached at the inclusion of the top 10 variables, beyond which the inclusion of additional variables did not meaningfully increase accuracy.

The datasets represented in Table 1 are those deemed by the RF classifier to have a measureable impact on classifier accuracy. Geophysical data included the thorium radiometric channel, EM and the RTP magnetics in addition to the DTM which ranked highest. Figure 3 shows clearly the occurrence of geometries in the EM data due to geology, in particular a NE trending

Rank	Variable	RF Score	10-Folds Acc
1	DTM	5.20	0.704
2	As	4.00	0.771
3	Th (radiometric)	3.40	0.782
4	Ta	3.30	0.815
5	Mg	3.10	0.838
6	Ti	3.10	0.848
7	Emz4	3.00	0.868
8	Ni	2.80	0.871
9	La	2.80	0.869
10	RTP	2.60	0.888

Table 1. Top 10 datasets ranked by RF. RF score is a measure of error variability produced by variable. 10 folds Acc was the accuracy achieved during 10 fold cross validation performed using the corresponding variable in addition to those ranked higher.

antiform pattern in the central southwest of the project. The DTM shows some geological control on topographic features (Figure 3) but is also likely serving as a proxy for position, hence the high ranking. From the geochemistry, long known immobile trace elements (e.g. Pearce & Norry, 1979; Maclean & Barrett, 1999) Ti and Ta featured prominently. Ta scored a 0.98 correlation with Nb, resulting in the omission of Nb prior to ranking through an objective process. While a subjective decision may have retained Nb due to greater abundance, the high correlation in this case indicates Ta performs equally well, making this, along with Ti a geochemically meaningful selection. Clear zonation can be seen in the Ta dataset (Figure 3) indicating a major domain change. As, Mg, Ni and La were also featured in the ranking.

CONCLUSIONS

Ranking via RF in this simulation has provided a list of datasets deemed most productive to classification. This ranking of datasets provides additional value in facilitating the prioritisation of datasets for any further interpretation. The fact that prominent geophysical mapping datasets and several elements that are well known lithological discriminators such as Ti and Ta (or equally Nb, with an r=0.98 correlation) are ranked highly lends confidence to the validity of these rankings for use in other geological interpretations outside of the machine learning space.

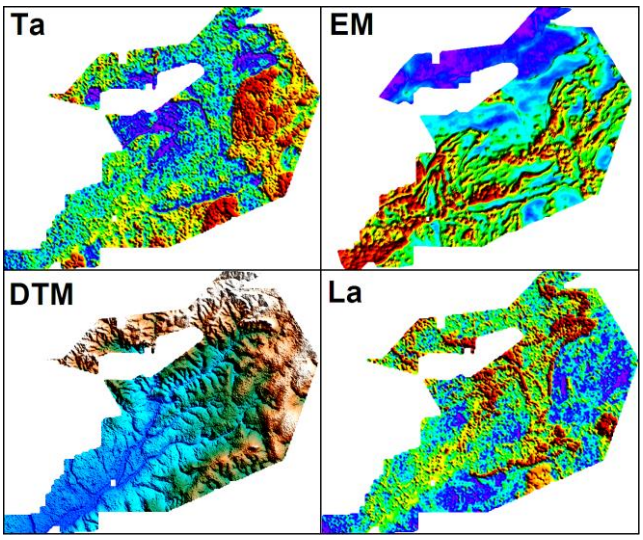

Figure 3. Examples of datasets ranked as important to RF constructing the classifier used in this study (Ta, EM, La and the DTM). The extent of these datasets corresponds to project area shown in Figure 1.

The RTP magnetics dataset was ranked as necessary albeit of low importance while the first vertical derivative was redundant. This could be due to the similarity in magnetic signature across multiple lithologies. Alternatively, the higher frequency magnetic data may be mapping sub-units or other textures within lithological domains and thus the magnitude of magnetic response cannot be diagnostic of lithology at the scale of this investigation. This is an interesting finding as magnetic data are commonly used as a primary mapping and interpretation tool. In this case, use of objective ranking would caution the user that while this dataset may be extremely useful for structural and textural mapping, it may be unreliable for distinguishing one lithological unit from another. It is important to note that this could also be an indication that available outcrop poorly expresses the variability of magnetisation in the project area. In this case, the usefulness of magnetics will manifest with the use of more spatially representative training data. Nevertheless, these ranking indicate that attempting to use magnetics to propagate beneath cover, the rock types seen in outcrop, would not be productive in this scenario.

Several other elements were identified as important during ranking. Geochemists can rationalise their use in most cases. For example, the La dataset (Figure 3) shows a strong trend in the centre northeast of the project which company geologists (Ireland, pers coms, 2016) identify as a monazite trend within a unit of Roan group sediments. Additionally, company geochemists have used As and Mg for subdivision of mafic packages and partitioning of talc rich rock units respectively. This further demonstrates that RF produces rankings that are geologically meaningful.

We have demonstrated that in addition to a lithological prediction, as a bi-product, RF provides geoscientists with a means of identifying and ordering the datasets most relevant to mapping of a project. The high importance of well subscribed mapping datasets in RF ranking, lends confidence that other datasets deemed important by RF are geologically sound albeit in a project specific context. In this case, as noted above, well-established datasets such as topography or high field strength elements are simple to rationalise, while others are idiosyncratic to the Trident project, such as the use of La for mapping monazite. Geochemical knowledge of the Trident project explains the influence other elements consistent with RF rankings.

Our findings make an important demonstration of a method capable of providing a rapid means of prioritising data for interpretation. We do not suggest that a skilled geoscientist cannot perform this task; rather, this method is able to do so in an objective manner and in the absence of geophysical or geochemical expertise. Additionally, this method reduced the number of datasets required for any first-pass exercise from 59 to 10, giving an interpreter a much simpler starting space to begin work. This is of particular value in the early stages of a project, where a robust understanding of the available data is under-developed. Ranking via RF can assist in the development of such an understanding.

ACKNOWLEDGMENTS

This research was conducted as part of the ARC Research Hub for Transforming the Mining Value Chain (project number IH130200004). Stephen Kuhn is supported through a Research Training Scholarship through the University of Tasmania. We would like to thank First Quantum Minerals Ltd. for access to data and permission to publish this study. RF and Pythagorean tree visualisations were performed using the Orange software package (Demsar et. al., 2012). Data pre-processing and plotting was performed using Geosoft, Oasis Montaj and ESRI, ArcGIS .

REFERENCES

Beck, F., Burch, M., Munz, T., Di Silvestro, L. and Weiskopf, D., Generalized Pythagoras Trees for Visualizing Hierarchies. 9th International Conference on Information Visualisation Theory and Applications, 17 - 28.

Breiman, L., 2001, Random Forests. Machine Learning 45, 5-32.

Capistrant, P.L., Hitzman, M.W., Kelly, N.M., Kuiper, Y., Wood, D., Williams, G., Zimba, M., Jack, D., and Stein, H., 2015, Geology of the Enterprise Hydrothermal Nickel Deposit, North-Western Province, Zambia. Economic Geology 110, 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. Computers&Geosciences 63, 22 - 33.

Cracknell, M.J., Reading A.M., McNeill A.W., 2014, Mapping Geology and Volcanic-Hosted Massive Sulfide Alteration in the Hellyer–Mt Charter Region, Tasmania, Using Random Forests™ 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, 2349−53.

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

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.

MacLean, W.H. and Barrett, T.J., 1993, Lithogeochemical Techniques Using Immobile Elements. Journal of Geochemical Exploration 48, 109-33.

Pearce, J.A. and Norry, M.J., 1979, Petrogenetic Implications of Ti, Zr, Y, and Nb Variations in Volcanic Rocks. Contributions to Mineralogy and Petrology 69, 33-47.