Can machine learning help us map geology beneath cover?

Rhys S. Davies, McLean Trott, Mike Dentith, David I. Groves & Allan Trench

Why predict bedrock geology?

Over the last decade, the exploration industry has moved from creating wealth to destroying wealth

This trend is attributed to an over-focus on near-mine brownfields exploration, providing a long-term diminishing rate of return as resources are depleted

Targeting below deeper, more complex cover is recognised as key to opening up new search-spaces and re-invigorating exploration success through significant new greenfields discoveries

An accurate bedrock geology map is critical to supporting exploration targeting efforts beneath cover

What is Random Forests?

The Random Forests machine learning algorithm is a supervised ensemble classification algorithm (Breiman, 2001). It is an extension of the decision tree approach, where leaves represent classes and branches represent features that lead to those classifications. Decision trees are typically built in a top-down manner; choosing a feature at each step that best splits the sample set.

Rather than producing a single tree, Random Forests applies an ensemble method by randomly selecting subsets of the training data (known as bootstrap aggregating or bagging) to build multiple decision trees (Figure 1). The output of the Random Forest classifier is the class selected by largest number of individual trees. The benefit of an ensemble approach is that it helps to avoid overfitting of training data, a recognised issue for individual decision trees, typically improving classification of non-training data. This approach also generates a measure of confidence in the classification; based on the percentage of individual trees that agree with an assigned class.

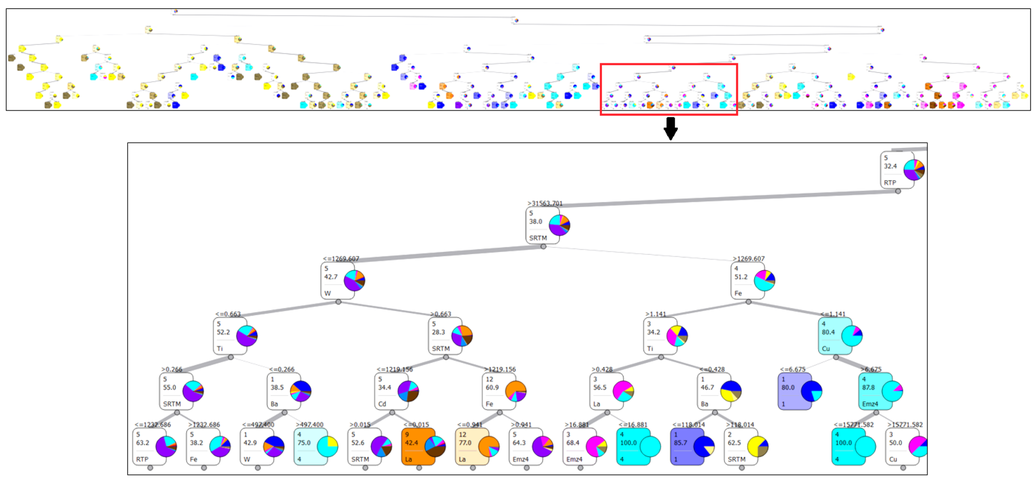

Figure 1. *Example of Random Forest decision tree. Source: Kuhn et al. (2016).*

What approach did we take?

The existing surface geology map, produced by the GSWA, was randomly sampled to select 15,000 training samples for each of 5 broad Archean geological classes. This produced a dataset of 75,000 samples, equivalent 5.5% of the full dataset of 1.34 million samples. These data were used as a training set for the application of Random Forests to predict the geology of residual material. The features used in this step (Table 1a) included geophysical and remote sensing data (magnetics, gravity, radiometrics, ASTER).

In this step, we used the trained Random Forest to classify ≈85,000 new samples, mapped as residual material. For this step, the classification of mapped Archean units was accurate in 98.8% of instances. This step was conducted in an effort to increase the total number and spatial coverage of data for training a second analysis, so as to improve the prediction results for areas beneath cover.

	Splits			Candidates			
Feature	Level 0	Level 1	Level 2	Level 0	Level 1	Level 2	Importance
40m_Mag1VD	17	30	58	17	46	107	2.194
40m_MagRTP	21	37	51	23	55	105	2.071
ASTER_Opq	21	27	34	31	51	105	1.531
80m_RadTC	11	23	43	22	49	109	1.364
80m_RadTh	12	19	49	31	53	120	1.154
400m_GravRTP	8	18	62	34	50	121	1.108
80m_RadU	8	12	23	25	55	92	0.788
80m_RadK	1	13	43	27	66	116	0.605
ASTER_FIs1	1	10	12	27	66	104	0.304
ASTER_FIs2	0	6	14	35	49	107	0.253
ASTER_FIs3	0	5	7	28	60	114	0.145

	Splits			Candidates			
Feature	Level 0	Level 1	Level 2	Level 0	Level 1	Level 2	Importance
40m_MagRTP	38	49	111	38	90	205	2.086
400m_GravRTP	40	69	118	53	104	209	1.983
40m_Mag1VD	22	53	92	51	97	194	1.452

Table 1a & b. *Feature importance for steps 1 and 2 (levels 0 to 2).*

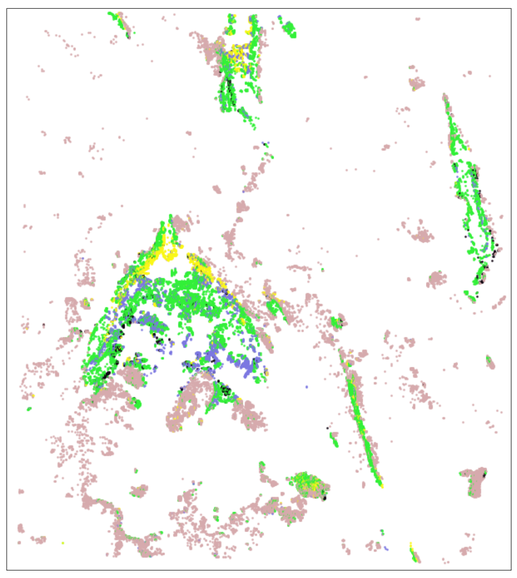

Figure 2. *Bedrock geology classification results from step 1, classifying geology of residual material.*

	Predicted Class (%)				
Actual Class (%)	Ag	As	Ab	Au	Ac
Ag	**93**	3	10	1	0
As	3	**91**	5	1	0
Ab	4	5	**68**	5	19
Au	0	1	12	**91**	0
Ac	0	0	5	2	**81**

Table 2. *Confusion matrix for step 2. A large proportion of mafic rocks were incorrectly classified as granitic or ultramafic rocks, likey leading to an over-abundance of these classes.*

A second analysis step was then conducted. The results of the first step were used as training data. Out of the mapped outcrop and classified residual material, approximately 32,000 training samples were randomly selected for each of the 6 geological classes. This produced a dataset totalling 160,000 samples, equivalent 12% of the full dataset of 1.34 million samples. These data were then used as a training set for the application of Random Forests to predict the geology beneath transported cover. The features used in this step (Table 1b) included only ground-penetrating geophysical data (magnetics and gravity).

Here, we used the trained Random Forest to classify ≈1.18 million new samples. These samples represented locations in the study area beneath transported cover. Classification of mapped Archean and residual units was accurate in ≈88.1% of instances. A confusion matrix (Table 2) is used to illustrate where misclassification occurred, showing that Random Forests performed well in the classification of granite vs greenstone units, but was prone to misclassification when differentiating between each of the separate greenstone units (Figure 3).

A measure of feature importance is built into Random Forest, showing us how effective the use of a given feature was in producing a decisive split between classes. Table 1a presents feature importance for step 1 and Table 1b for step 2).

References

Breiman L., 2001. Random Forrests, Machine Learning, v. 45, p. 5-32.

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, v. 61:2, p. 287-304.

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, v. 63, p. 22 – 33.

Davies, R. S., Groves, D. I., Trench, A., Sykes, J. P., Standing, J. G., 2018. Entering an immature exploration search space: Assessment of the potential orogenic gold endowment of the Sandstone Greenstone Belt, Yilgarn Craton, by application of Zipf's law and comparison with the adjacent Agnew Goldfield. Ore Geology Reviews , v. 94, p. 326–350.

Davies, R. S., Ryan, D., Groves, D. I., Trench, A., Sykes, J. P., Standing, J. G., Jia, C., and Robertson, W., 2017, Sandstone Goldfield; in Phillips, G N (ed), 2017. Australian Ore Deposits, 864 p (The Australasian Institute of Mining and Metallurgy: Melbourne).

Kuhn, S., Cracknell, M. J., Reading, A. M., 2018. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. Geophysics, v. 83:4, p. 183-193.

Kuhn, S., Reading, A. M., Cracknell, M. J., 2016. Geological Mapping in the Central African Copper Belt: How Would a Machine Do It? CODES, UTAS.

What did the results look like?

In this analysis, Random Forest successfully delineated the extents of the Archean Sandstone Greenstone Belt and surrounding minor greenstone belts. However, the accuracy of prediction within the belts remains poor, with the analysis struggling to differentiate between mafic, ultramafic and various sedimentary rocks (Figure 3).

Through the use of multiple decision trees, Random Forests calculates the probability of each class for each instance. This calculated 'Information Entropy' can be used to quantify uncertainty (Figure 4), and therefore used to focus future data collection during exploration.

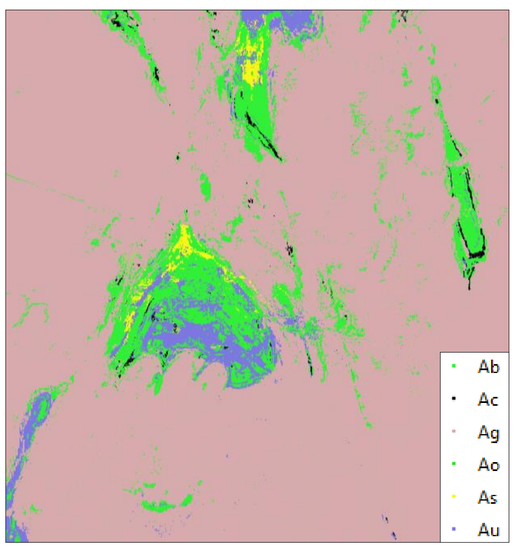

Figure 3. *Bedrock geology classification results from step 2, classifying geology beneath transported cover.*

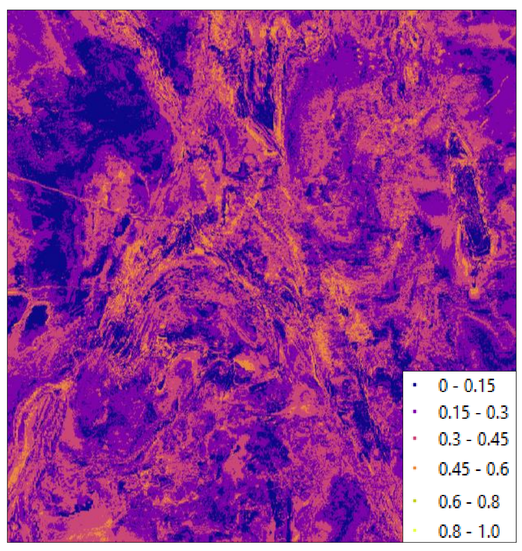

Figure 4. *Bedrock geology classification entropy scores from step 2 (blue = low, yellow = high).*

Where will we try to take this next?

This research remains very much at a trial stage, with several areas recognised as promising for improving overall analysis outcome, particularly with regards to differentiating between greenstone units.

To determine an optimum analysis method, we are in the process of:

- Testing multiple different training datasets; including outcrop mapping, existing bedrock interpretations, and geological logging from drillholes
- Incorporating multiple geophysical datasets and their derivatives as additional features
- Making use of cloud computing to conduct higher resolution analyses; by increasing the resolution of the study, it is expected that less well-represented lithology classes will be represented by a larger quantity of training data, potentially improving the overall analysis outcome
- Analysing a larger study area; again, this may help to increase the quantity of training data for less well-represented classes
- Trialing other supervised machine learning algorithms, such as neural networks support vector machines, etc.
- Further tuning for hyperparameter optimization
- Conducting unsupervised classification of each individual lithology class (as predicted by Random Forests); this allows us to explore each class for subtle differences, compare average geophysical responses between different clusters, and perhaps highlight incorrectly assigned clusters

Once a suitable bedrock geology product has been generated, the intention is to incorporate this layer into a prospectivity analysis to support exploration targeting efforts.
To conduct such an analysis, this bedrock geology layer would be combined with other layers (such as structure and soil geochemistry) that are considered relevant to targeting for mineralisation in the region.