Lithological mapping via Random Forests: Information Entropy as a proxy for inaccuracy

Stephen Kuhn*
*Sch. of Physical Sci. (Earth Sciences)
and CODES Centre of Excellence
University of Tasmania
stephen.kuhn@utas.edu.au*

Matthew J. Cracknell
*Sch. of Physical Sci. (Earth Sciences
and ARC ITR Hub for Transforming
the Mining Value Chain, UTAS
m.j.cracknell@utas.edu.au*

Anya M. Reading
*Sch. of Physical Sci. (Earth Sciences)
and CODES Centre of Excellence
University of Tasmania
anya.reading@utas.edu.au*

**Presenting Author*

SUMMARY

Machine Learning Algorithms (MLA) can be an effective means of lithological classification. The Random Forests^{TM} (RF) supervised classification approach allows prediction of lithology from disparate geophysical, geochemical and remote sensing data. In this study, we examine the relationship between prediction accuracy and information entropy (H). Data were processed in accordance with industry best practice and input selection was optimised using RF. Using a training set containing 1.4% of available pixels, we produced a classified lithology map with an overall accuracy of 76% with regards to mapped geology. In addition, we produced a class membership probability for each pixel, a precursor to defining the ultimate class designation at each pixel. H was calculated at each pixel from output class membership probabilities; and in this context provides a measure of the state of disorder for each. H was normalised with 0-1 representing the minimum to maximum possible H for each pixel.

H equal to 1 at a pixel represents an equal probability of all candidate classes occurring, whereas H equal to 0 describes a 100% probability of single class occurring. In this study, we demonstrate that there is a significant difference in the distribution of H between correctly and incorrectly classified pixels. The median H of incorrectly classified samples occurs above the 75% percentile of H for correctly classified samples. Conversely, both the mean and median H for correctly classified pixels occurs below the 25% percentile level for incorrectly classified samples.

This information can be used to determine the well-defined transition range in H, above which classification is likely to be inaccurate. Using this approach, a geoscientist can produce a lithological map, a quantifiable measure of uncertainty and a quantifiable transition range above which they are likely to encounter incorrect classification, avoiding wasted expense in targeting based on an incorrect model.

Key words: Supervised classification, Random Forests, Information Entropy, Lithological mapping

INTRODUCTION

MLAs are seeing increasing use as a means of applying data-mining principles for the benefit of geological applications. RF (Breiman, 2001) is an extension of the classification and regression tree method, involving the creation of an ensemble, or forest of decision trees, each voting towards the ultimate output. Randomness is introduced via bootstrap aggregation as a means of selecting the data to be used for each tree. Additionally, input variables used to split each node of a tree are selected at random. Each node is split to maximise improvement in homogeneity of the child node relative to the parent. RF has been shown to achieve similar or better accuracy than other classification algorithms while having the advantage of being resistant to over-fitting. It is also easier to use than many other options (e.g. Cracknell & Reading 2013, Hastie et. al., 2009). This makes RF an ideal choice for geoscientists as computational expertise is not a limiting factor in wider adoption of the method.

RF has been increasingly applied to the problem of lithological classification. Waske et. al. (2009) compared RF and another popular machine learning algorithm: Support Vector Machines (SVM); to standard classifiers in the context of mapping using hyperspectral imagery. They concluded that both methods achieved significantly more accurate results than traditional classifiers. While in that instance, SVMs marginally outperformed RF, it was noted by the authors that RF remained an attractive option due to high accuracy and relative ease of use. Cracknell & Reading (2013) assessed RF and SVM for lithology mapping and the identification of lithological contacts and zones of structural complexity. They demonstrated that RF, in addition to excellent overall performance, produced more usable outputs. Unlike SVM, high uncertainty was more strongly associated with incorrectly classified samples. A further, study by Cracknell and Reading (2014) compared RF with four other MLAs: SVM, Naïve Bayes, k-Nearest Neighbours and Artificial Neural Networks; as applied to lithological mapping. In their study, RF marginally outperformed other MLAs. While there were only small differences in accuracy, RF was able to produce these results with simpler input parameters and at less computational cost than other algorithms evaluated.

Various metrics are increasingly being used as a means of defining the uncertainty associated with predictive modelling. Information Entropy (H) (Shannon, 1948), defined in equation 1, has been used to great effect in a "per-voxel" setting to demonstrate how uncertainty is distributed spatially (Wellmann & Regenauer-Leib, 2012). In this study, we will demonstrate that H can be used as a

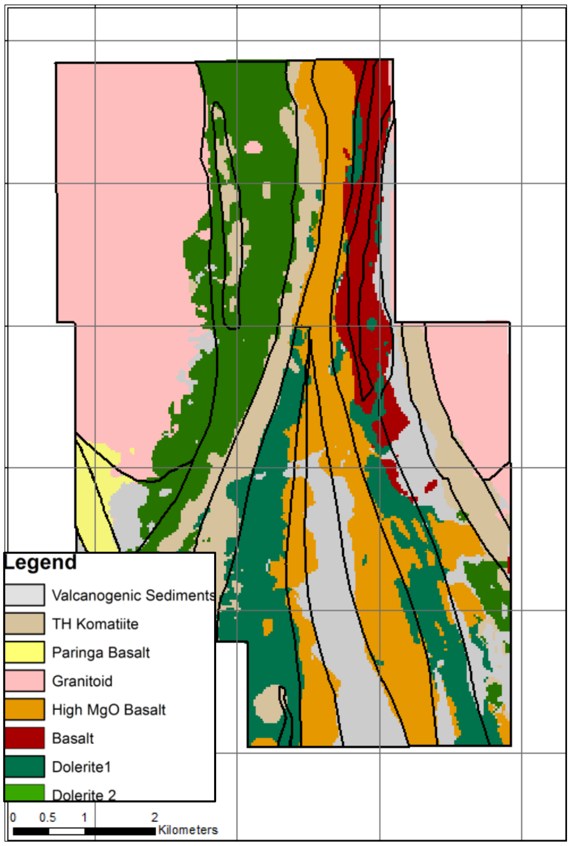

Figure 1. Lithology as predicted by RF. Graticule spacing is equal to 2000 m. Black lines indicate lithological boundaries in original interpreted geological map.

value adding product of the RF classification process, providing not only a measure of uncertainty, but provide a cut-off range above which classification is likely to be inaccurate.

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

Equation 1. Information Entropy, where p_i=class membership probability at location i, n= number of candidate classes and K is a positive constant determined by choice of unit of measure.

In this study, we use RF to classify pixels by lithology in an orogenic gold prospective setting, from the Eastern Gold Fields, Australia. The area comprises a NS striking, steeply dipping package of mafic and sedimentary units impinged between larger granitoid bodies to the east and west. Our primary objective is to calculate H for classified pixels and analyse distribution within correctly and incorrectly classified results. We aim to assess whether there is a significant difference in H between the correct and incorrect sets and a threshold by which they can be separated. As such, the experiment is set up such that the null hypothesis is as follows: there is no difference between the information entropy attributed to samples classified as correct or incorrect.

METHOD AND RESULTS

RF Classification

RF was used to classify 55,995 samples into eight lithological classes. Eight variables were measured at each sample point (reduced from 16 using RF to eliminate irrelevant inputs). These variables were as follows: Bouguer Anomaly and its first vertical derivative; Reduced to Pole Total Magnetic Intensity and its first vertical derivative, airborne radiometrics (K, U and Th) and a digital terrain model.

Each sample point represents a 30 x 30 m pixel in the gridded datasets. Approximately 1.4% of data, selected via spatially stratified random sampling were used independently to train a RF classifier. The trained classifier subsequently predicted a lithological class for each of the remaining ≈98.6% of pixels into one of the eight geological classes: Granite, Dolerites (type 1 and 2), Sediments, Basalt, High MgO Basalt, Paringa Basalt or Tripod Hill Komatiite. A lithological map was produced in addition to class membership probabilities for each class (Figure 1). Results of the classification were validated with respect to known/mapped geology for each pixel, with an overall accuracy of approximately 76% (see Figure 3).

Calculation and analysis of H

Class membership probabilities (Figure 2) calculated by RF (probability of a pixel belonging to each available class) were used to calculate H. The properties of H are such that a value of 0 corresponds to a 100% probability of 1 class occurring and a value of 1 corresponds to an equal probability of all represented classes being present.

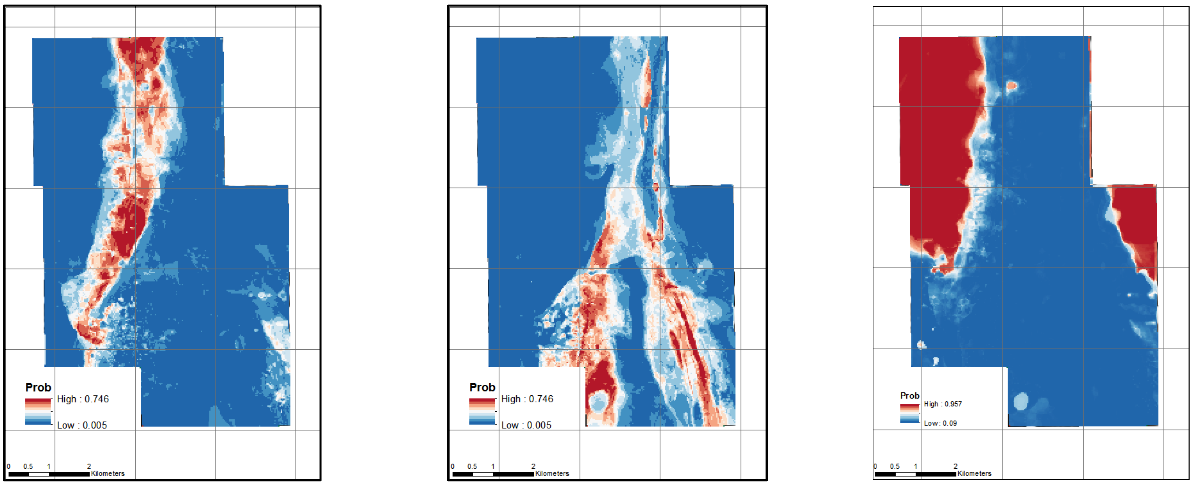

Figure 2. Examples of spatial distribution of class membership probabilities generated by RF. Left: Dolerite type 2. Centre: Dolerite type 1. Right: Granitoid.

H in its general form preserves monotonicity such that an increase in the number of candidate classes results in higher H. For the purpose of this study, H has been normalised to account for number of candidate classes, such that H assigned to each pixel represents, on a scale of 0-1, the range of minimum to maximum possible H for that pixel. As such, all pixels are comparable in regards to how close they each internally approach their minimum or maximum possible H. For example a pixel with two equally probable classes and a pixel with four equally probable classes shall both describe H equal to 1. The spatial distribution of H is shown in figure 3.

H for all samples was split into two groups. One containing samples classified correctly and the other containing samples classified incorrectly in order to analyse distribution of H both within and between these two groups, and test the null hypothesis. Figure 3 shows the interclass vs intra-class distribution of H it can be observed qualitatively that there is a significant difference between the values of entropy represented by the two sets. For the benefit of mathematical completeness, a non-parametric Wilcoxon signed-rank test was performed, producing a p value of $2.2e^{-16}$, being the minimum software value and effectively approaching zero, demonstrating with near certainty that there is a statistically significant difference between groups.

In further analysis, it can be seen that both the mean (H = 0.57) and median (H = 0.68) value for correctly classified samples lie below the 25% percentile value of incorrectly classified samples (H = 0.7). Conversely the median (H = 0.77) value of incorrectly classified samples lie above the 75% percentile value (H = 0.76) for correctly classified samples while the mean (H = 0.75) is approximately equal to the 75% percentile value (Figure 4).

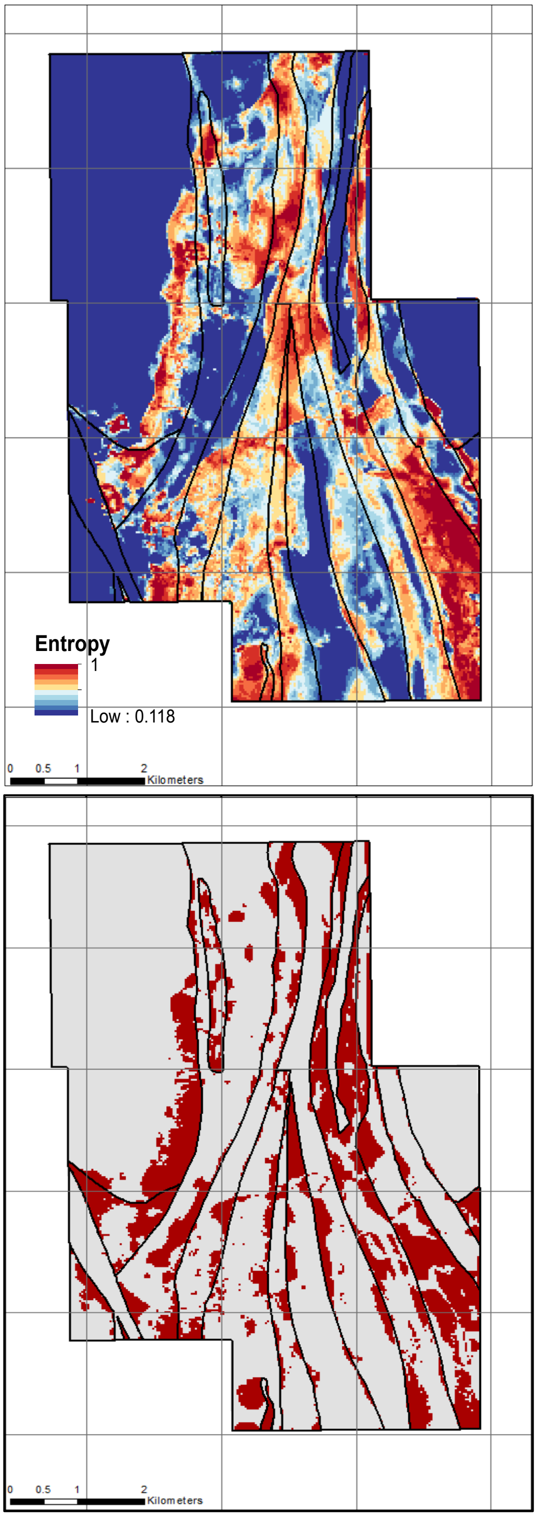

Figure 3. Top: Spatial distribution of H. Bottom: Accurate (grey) versus inaccurate (red) pixels as compared to interpreted geological map Black lines indicate boundaries defined in original interpreted geological map. Graticule spacing is equal to 2000 m.

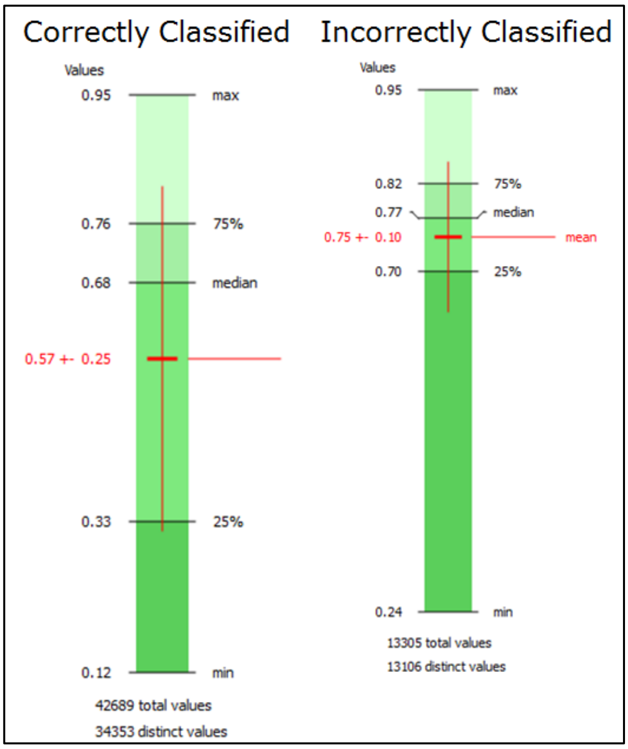

Figure 4. Within and between group distributions of H For correctly and incorrectly classified groups. Vertical red line indicates 1 standard deviation from the mean.

Additionally, we examined the relationship between the proportion of correctly classified samples which at a given threshold exhibited values of H greater than that threshold and conversely, the proportion of incorrectly classified samples that for a given threshold exhibited a value of H less than that threshold. This method, as used by Cracknell & Reading (2013) produces a curve for each of the two groups, the intersection of which defines a point representing the optimal trade-off between preserving correctly classified samples whilst eliminating incorrectly classified samples. This occurred at H= 0.76 (see Figure 5).

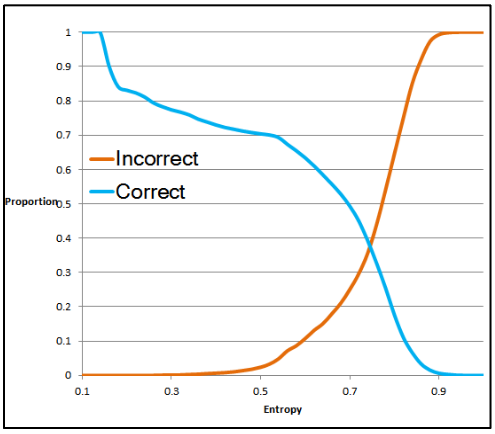

Figure 5. Blue: The proportion of correctly classified samples with H greater than a given level. Orange: The proportion of incorrectly classified pixels with H less than a given threshold

CONCLUSIONS

From our results, it is clear that there is both a statistically significant and qualitatively observable difference in H between correctly and incorrectly classified pixels. There is significant overlap of H between correctly and incorrectly classified samples however some clear ranges can be defined to assist geoscientists in validating their mapping efforts. Namely, a value of H of 0.76 was discovered to be the optimal cut-off for H that preserved the maximum number of correctly classified pixels whilst eliminating the maximum number of those which were misclassified.

Analysis of the distribution of H within and between correctly and incorrectly classified samples can provide useful insight to a geoscientist in regards to the level at which they may be including incorrectly mapped pixels or excluding correctly classified pixels into their final outputs when modulating a threshold value for H. The implication being that the first scenario could be far more costly in that expense of additional, unnecessary map validation is significantly less expensive than drilling a null target based on mapping falsely assumed to be correct. By understanding the rate at which these errors occur at various thresholds, a geoscientist can make a more informed decision when trying to mitigate risk.

A significant limitation of this study is in the accurate definition of correct versus incorrect, being validated using an interpreted geological map which in itself may contain errors or more likely, an oversimplification of geology. Optimally, the deployment of these methods would be of maximum value with H thresholds calculated in relation to accurate observed geology and subsequently applied to poorly mapped, un-mapped or undercover areas. This requires the assumption that the relationship between H and accuracy will be retained, which is a reasonable ± small error factor, provided the method is applied intelligently to known and unknown regions within a geological domain or similarly defined congruent region. Provided these criteria are met, this study demonstrates H to be an effective means of thresholding for accurate versus inaccurate Lithology predictions. This can be used to better target expenditure towards regions which have not been adequately classified and thus require further mapping. Simultaneously, it aids in risk mitigation through reducing wasted expense in the form of drilling or other follow up work based on inaccurate models of the Earth

ACKNOWLEDGMENTS

This research was conducted in association with the ARC Research Hub for Transforming the Mining Value Chain (project number IH130200004). Stephen Kuhn is supported through an Australian Postgraduate Award Scholarship through the University of Tasmania. We would like to thank Gold Fields Ltd. for access to data for the purpose of this study. The Orange software package (Demsar et. al., 2013) was used for RF classification. Pre-processing and plotting was performed using Geosoft Oasis Montaj and Esri ArcGIS.

REFERENCES

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

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.

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**, WB113 - WB26.

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.

Shannon, 1948, A Mathematical Theory of Communication. *Bell Systems Technical Journal* **27**, 379-423.

Waske, B., Benediktsson, J.A., Árnason, K., and Sveinsson, J.R., 2009, Mapping of Hyperspectral Aviris Data Using Machine-Learning Algorithms. *Canadian Journal of Remote Sensing* **35**, 106-16.

Wellmann, J.F. and Regenauer-Lieb, K., 2012, Uncertainties Have a Meaning: Information Entropy as a Quality Measure for 3-D Geological Models. *Tectonophysics* **526–529**, 207-16.