Supervised and unsupervised classification of near-mine soil Geochemistry and Geophysics data

Matthew J. Cracknell
ARC Centre of Excellence in Ore Deposits (CODES), University of Tasmania
Private Bag 126, Hobart, Tasmania, 7001
M.J.Cracknell@utas.edu.au

Anya M. Reading
School of Earth Sciences, University of Tasmania
Private Bag 79, Hobart, Tasmania, 7001
Anya.Reading@utas.edu.au

Andrew W. McNeill
Mineral Resources Tasmania, Dept. Infrastructure, Energy & Resources
PO Box 56, Rosny Park, Tasmania, 7018
Andrew.McNeill@dier.tas.gov.au

SUMMARY

Remotely sensed geoscience data can assist detailed geological field mapping in areas of thick vegetation and poor outcrop. However, the potentially high dimensionality of these data makes it difficult to visually interpret and fully comprehend. Machine learning algorithms provide an efficient semi-automated means of recognising and identifying patterns in data. We use Random Forests for supervised classification of geologic units from airborne geophysical and soil geochemical data in the economically significant Hellyer - Mt Charter region of western Tasmania. A backward-recursive variable selection method is used to select the most relevant and useful data for this problem. This reduces computation cost and enhances interpretation of results without significantly affecting prediction accuracy. Random Forests generates accurate predictions of the spatial distribution of surface geologic units from these data. An example is provided regarding the use of Self-Organising Maps, an unsupervised clustering algorithm, to identify distinct but spatially contiguous clusters within a geologic unit. By visualising cluster spatial distribution and identifying key variable contributions to cluster differences, we interpret the geological significance of intra-class variability.

Key words: Supervised classification, Unsupervised clustering, Geophysics Geochemistry, Tasmania.

INTRODUCTION

The Hellyer - Mt Charter region hosts three known economic Volcanic Hosted Massive Sulphide (VHMS) Pb-Zn ore bodies. The dominant rocks in this region are a sequence of Cambrian marine calc-alkaline mafic to felsic volcanics and volcaniclastics known as the Que-Hellyer Volcanics (Waters & Wallace 1992). Although there is little Quaternary glacial cover in the area, detailed field mapping is a challenging prospect due to poor outcrop, dense rainforest vegetation and thick soil profiles (Corbett & Komyshan 1989). Thus, additional information on the spatial distribution of surface geology will be of significant value to mineral explorers. Mineral exploration in the region dates from the early 1970s has generated a large amount of geophysical and geochemical data. Nevertheless, it is difficult for geologists to analyse and interpret this data especially where geological understanding comes from complex variable interactions. We use Random Forests for the supervised classification of surface geology based on training data obtained from a pre-existing geologic map. Self-Organising Maps are employed to find spatially contiguous clusters within individual geologic units using key input variables.

Machine learning algorithms provide users with robust discrimination capabilities for supervised and unsupervised classification problems especially where the dimensionality of the input variables is high and where the relationships between data and outputs are poorly understood. They offer opportunities for data users to objectively identify patterns in multi-dimensional input variables using an inductive approach (Hastie *et al.* 2009). The Random Forests algorithm (Breiman 2001) constructs an ensemble of Classification and Regression Tree (CART) classifiers from training data using bagging and random variable selection. The Gini index measures class heterogeneity and is used by Random Forests to determine a “best split” threshold at each node of a tree. Predictions are based on a majority vote cast by the ensemble of CART classifiers (Breiman 2001, Hastie *et al.* 2009). Random Forests is effective at predicting spatially distributed geological classes using multisource and high dimensional remote sensing data (Waske *et al.* 2009; Cracknell and Reading, under review). The results of these studies show Random Forests performed as well or significantly better than standard classifiers (e.g. Maximum Likelihood) and other machine learning strategies (e.g. Artificial Neural Networks and Support Vector Machines). Moreover, Random Forests is efficient, insensitive to noisy data and results in good performance when faced with limited training data.

Self-Organising Maps (SOM) (Kohonen 1982) are information-driven data mining tools that have the ability to highlight subtle relationships within high dimensional and seemingly disparate datasets (Fraser & Dickson 2007). SOM is an unsupervised clustering algorithm proven to be useful and efficient for exploring high dimensional geoscience data (e.g., Bierlein *et al.* 2008, de Carvalho Carneiro *et al.* 2012). SOM treats each sample as an *n*-dimensional vector in variable space. Using vector quantisation and vector similarity measures (i.e. Euclidian distances) SOM segments data into a number of distinct clusters via an iterative two-stage process. Initially, input samples closest to randomly placed seed-nodes are identified. Seed-nodes are then “trained” such that their values are adjusted in order to align closely to clusters within the input samples. This generates a mapping of inputs to trained seed nodes (node-vectors) onto a 2D space. The topology between node-vectors (clusters) is preserved such that those close in nD space maintain their relative proximities on the 2D map (Bierlein *et al.* 2008).

The main aims of this study are to: 1) demonstrate the feasibility of Random Forests for remote sensing geological supervised classification applications; and 2) implement SOM to identify intra-class clusters and interpret their geological significance by visualising differences in key geophysical and geochemical input variables. Data preprocessing, variable selection, implementation of Random Forests and SOM, and visualisation of their outputs were conducted using the R Programming language (Comprehensive R Archive Network, http://cran.r-project.org/).

Geology and Data

The geology of the Hellyer - Mt Charter area has been divided in to 21 distinct lithological units (Figure 1). The most economically important units are within the Mixed Sequence of the Que-Hellyer Volcanics (QHV). The QHV contains four stratigraphic subdivisions. At its base lies the Lower Basalt (LB), exhibiting highly variable thickness. Conformably overlying LB is a unit of feldspar-phyric andesites (Afp) that underlies the majority of VHMS mineralisation. The Mixed Sequence, which is the host horizon for economic mineralisation, comprises strongly altered rocks (HA), polymictic volcaniclastics (Y), andesites (A) and dacites with minor basalts (D). Lateral facies variations are rapid within the Mixed Sequence and Y may be intensely hydrothermally altered as can the margins of some bodies of D (Waters & Wallace 1992). The upper most unit in the QHV is a sequence of coherent pillowed to massive basaltic rocks called the Hellyer Basalt (HB). Conformably overlying and intercalated/intermixed with the HB is the black pyritic/carbonaceous Que River Shale (QRS). Within the study region are a series of NE trending normal faults linked by NW trending transfer faults of probable Cambrian age. These structures are interpreted from a combination of gravity and magnetic surveys, rapid changes in volcanic facies and variations in unit thickness (McNeill *et al.* 1998). Several regional scale Late Cambrian-Devonian faults dissect the study region accompanied by NNE plunging folds. Fold wavelengths vary from 2 km to 500 m (Corbett & Komyshan 1989).

Data initially included in this study consisted of 14 interpolated and levelled geophysical variables, 11 C-horizon soil geochemistry variables and 7 bands of Landsat ETM+ imagery. Where required, these data were resampled (using bilinear interpolation) to a resolution of 20 m. All data were cropped to an area determined by the spatial distribution of soil samples. Geophysical data were smoothed using a 5x5 mean convolution filter to reduce the effect of artefacts (e.g. high voltage powerlines). Variables with log-normal distributions were transformed to normal distribution using the natural logarithm. All data were scaled to zero mean and unit variance. Total Magnetic Intensity (TMI) was Reduced to Pole (RTP) using Aust. Geomagnetic Reference Field (http://www.ga.gov.au) parameter estimates. A regional magnetic gradient (Richardson 1994) was removed by linear trend surface subtraction. Gamma-Ray Spectrometry (GRS) data were corrected for negative values. GRS band ratios, K/Th, U/Th, and U^2/Th, were calculated and included. Five data layers were present in the $\log_{10}$ AEM apparent resistivity maps relating to different frequencies of data acquisition. Soil geochemistry data contain ~26,000 individual georeferenced samples collected by hand auger, to a depth of 1 m. These samples are assumed to relate to C-horizon of the soil profile. Samples were analysed for up to 11 elements: Cu, Pb, Zn, Ag, Au, Ba, As, Cr, Zr, Ti, and Ni. Mean element abundances (ppm) were calculated for samples collected within 5 m of each other. Au data was not used in this study as there were <200 non-missing values. Interpolation from irregularly spatially distributed samples to a uniform grid with 20 m cell resolution was carried out using ordinary kriging with fixed covariance parameters for a global neighbourhood. Covariance parameters were estimated by fitting parametric exponential or spherical models, using weighted least squares, to an isotropic empirical variogram for distances of 3 km or 5 km at 100 m intervals. Landsat ETM+ image data, unaffected by cloud cover, were supplied with Level 1 processing. Band ratios, useful for discriminate geological materials (e.g., Durning *et al.* 1998, Boettinger *et al.* 2008), were calculated.

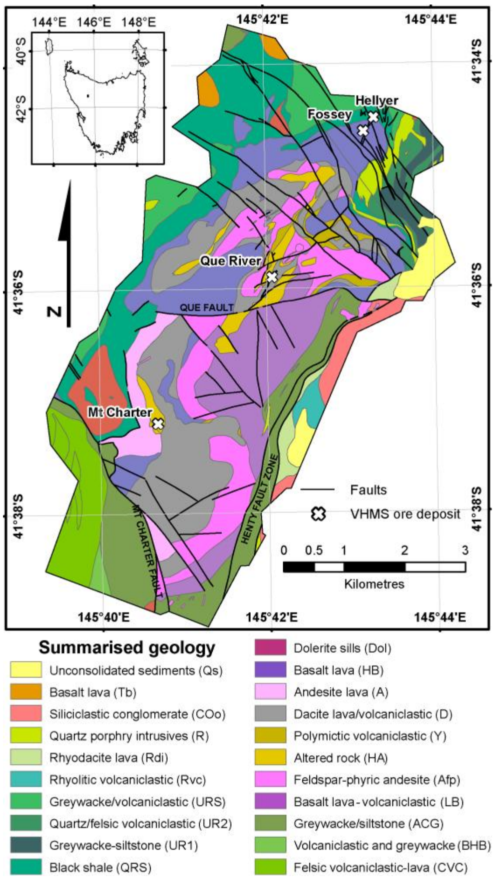
144°E 146°E 148°E
40°S
42°S
145°42'E
145°44'E
41°34'S
Hellyer
Fossey
Que River
QUE FAULT
N
41°36'S
Mt Charter
MT CHARTER FAULT
HENTY FAULT ZONE
Faults
VHMS ore deposit
41°38'S
0 0.5 1 2 3
Kilometres
145°40'E
145°42'E
145°44'E
Summarised geology
Unconsolidated sediments (Qs)
Basalt lava (Tb)
Siliciclastic conglomerate (COo)
Quartz porphry intrusives (R)
Rhyodacite lava (Rdi)
Rhyolitic volcaniclastic (Rvc)
Greywacke/volcaniclastic (URS)
Quartz/felsic volcaniclastic (UR2)
Greywacke-siltstone (UR1)
Black shale (QRS)
Dolerite sills (Dol)
Basalt lava (HB)
Andesite lava (A)
Dacite lava/volcaniclastic (D)
Polymictic volcaniclastic (Y)
Altered rock (HA)
Feldspar-phyric andesite (Afp)
Basalt lava - volcaniclastic (LB)
Greywacke/siltstone (ACG)
Volcaniclastic and greywacke (BHB)
Felsic volcaniclastic-lava (CVC)

Figure 1. Geology of Hellyer - Mt Charter study region, after Corbett and Komyshan (1989), Waters and Wallace (1992) and Richardson (1994)

METHOD AND RESULTS

Supervised classification requires prior information (training data) in order to construct a classification model. In this study, we randomly sampled 100 instances (pixels) for each of the 21 geologic classes. This reduced the potential for constructing a biased classifier resulting from a large difference in sample counts for the classes in the training data (Japkowicz & Stephen 2002). Spatially randomly sampled independent test dataset was used to evaluate Random Forest classification model performance. Test data contained an equal number of samples to the training data (i.e. 2100) resulting in class sample proportions approximately equivalent to class coverage over the study region.

Selection of non-redundant inputs was conducted by eliminating variables with mean correlation coefficients >0.8

associated with a large proportion of other data. A recursive (backward) variable selection method was then employed to find the minimum number of optimal inputs. Variables are ranked, using Random Forest importance measures, and recursively eliminated while estimating prediction accuracy. The minimum number of important variables was identified using a 0.015 threshold from the maximum mean accuracy. Mean accuracies were calculated using 10-fold cross-validation resampled over 10 iterations. Figure 2a plots cross-validation accuracy as a function of the 24 ranked variables after the removal of redundant data. Figure 2b presents the 11 selected variables in ranked order of importance. Selected variables contain geophysics data and soil geochemistry data.

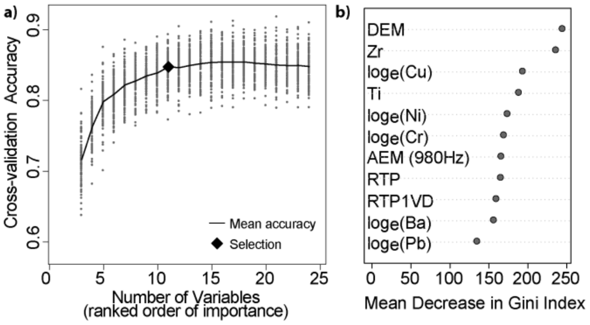

Figure 2. a) Random Forests variable selection cross-validation accuracies, and b) final list of selected variables in ranked order of relative importance

Random Forests requires selection of the number of variables to split (*mtry*) and the number of trees to construct (*trees*) to optimize performance. *Trees* was set to 500, while *mtry* values were varied and selected using maximum mean cross-validation accuracy. SOM was implemented with a 3x3 hexagonal topology using a neighbourhood radius equal to 2/3 of all variable ranges and learning rate decay from 0.05 to 0.01 over 100 iterations. A hierarchical (tree-based) clustering method was used to merge SOM derived clusters by iteratively combining clusters with similar properties until the desired number of clusters remain. Cluster spatial distribution was assessed in map form. The contributions of input variables to class sub-groups were examined by plotting frequency densities showing relative distributions of samples belonging to individual clusters.

The test data confusion matrix (not shown) has an overall accuracy of 0.784 ± 0.018 based on Exact 95% Confidence Limits. The majority of misclassifications occur between classes within the QHV. HA predictions generally misclassify D and to a lesser extent Afp. Y is equally misclassified as Afp, D and HB. Classes underlying or immediately above HA and Y and classes in the upper sequence of the QHV are misclassified as each other. Neither HA or Y are misclassified as each other despite their stratigraphic and spatial proximity. Due to strong basalt geochemical signatures, QRS is most often misclassified as HB. QRS is also misclassified as undifferentiated URS of the Southwell Sub-Group, which contains a large proportion of shale. Figure 3 provides a map geologic classes predicted by Ransom Forests, reclassified using a 3 x 3 majority convolution filter. Predictions generally conform to the interpreted extent of geologic classes in the reference map (see Figure 1). The inset in Figure 3 indicates the majority of misclassifications occur between spatially adjacent classes. These errors are concentrated in regions of complex faulting and thin and discontinuous Mixed Sequence units. Immediately to the north of Mt Charter, a zone of dense misclassifications is associated with the Mixed Sequence, especially where HB is mapped to underlie D and Afp is predicted as A, probably due to their similar geochemical signatures. QRS is difficult to correctly classify where it is mapped as a thin unit: in contact with HB and URS, and adjacent to dolerite west of Mt Charter.

Samples predicted as HB were grouped into four clusters. The hierarchical cluster dendrogram in Figure 4a indicates HB_1 and HB_3, and HB_2 and HB_4 are most similar. The map of the spatial distributions of intra-class clusters (Figure 4b) shows HB_4 is a thin and discontinuous layer along the lower margin HB, while HB_2 occurs stratigraphically above HB_4 north of the Que Fault. HB_1 is positioned within the hinge zone of an anticline north of the Que Fault and HB_3 occurs south of the Que Fault, coincident with the Mt Charter Pb soil anomaly and hinge zone of a syncline. Major differences between the four HB clusters occur within Ti, Cu, Cr and Pb variables (Figure 4c), with HB_1 and HB_3 exhibiting relatively higher Ti, Cu and Pb than HB_2 and HB_4.

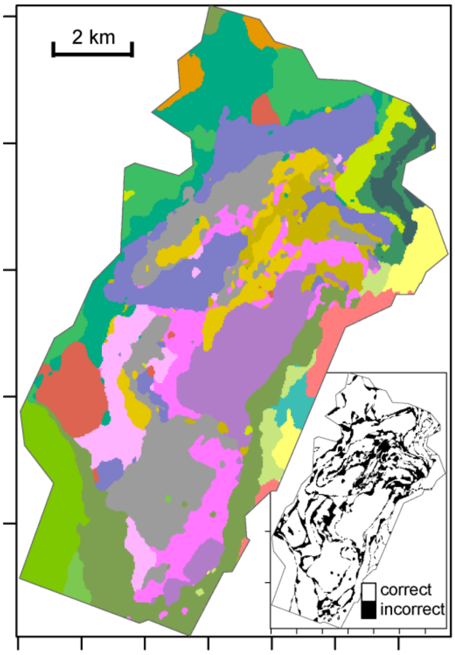

Figure 3. Geologic class predictions generated by Random Forests and spatial distribution of misclassified samples.

CONCLUSIONS

A backward-recursive variable selection method was used to identify and select the most important variables for the supervised classification of surface geologic units in the Hellyer - Mt Charter region using Random Forests. A significant reduction in the number of redundant and irrelevant variables improves the interpretability of results and reduces computation cost without significantly affecting classification accuracy. Random Forest categorical predictions of geological units generally conform to the reference geological map. However, the method we used to obtain training data assumes prior knowledge of the general extent of geological classes, information which may not be available in other applications. Random Forest predictions indicate the presence of small areas of potentially unmapped VHMS host horizons (i.e. HA and Y) south of the Que Fault and east of Que River,

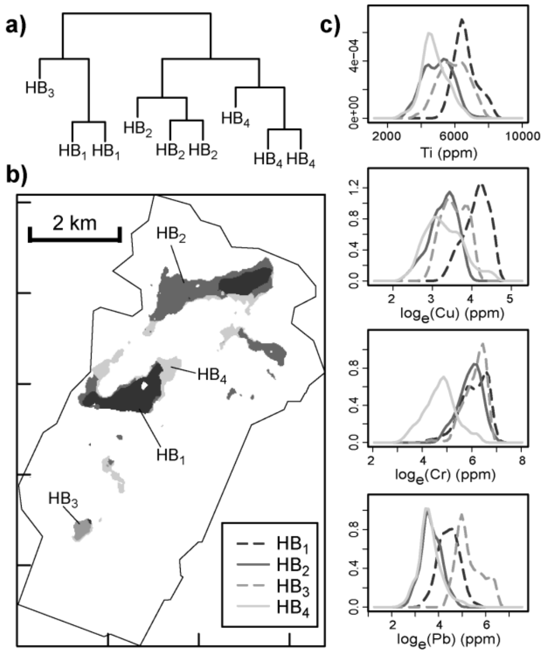

Figure 4. Self-Organising Map unsupervised clustering for Hellyer Basalt class predicted by Random Forests, a) hierarchical partitioning of clusters, b) spatial distribution and c) frequency density estimations of key variables

which may be linked to subsurface mineralisation. Furthermore, Random Forest misclassifications indicate the highly variable and mixed geochemical signature of the Que-Hellyer Volcanics. We used a hierarchical tree clustering method for visualising the appropriate number of intra-class clusters and merging similar Self-Organising Map node-vectors. We have identified four distinct intra-class groups within the Hellyer Basalt based on key geophysical and geochemical variables. The relative low base metal content and spatial distribution of clusters HB$_2$ and HB$_4$, suggests primary effusive origins. In contrast, HB$_1$ and HB$_3$ are likely to be associated with synvolcanic base metal mineralisation. Random Forests offers geologists an opportunity to accurately and objectively predict geology from multisource and high dimensional geoscience data. We show that by using Self-Organising Maps, geologically relevant and spatially contiguous clusters can be inferred within predicted classes. Visualising the spatial distribution and similarities and dissimilarities between clusters provides clear interpretations of their geological significance.

ACKNOWLEDGMENTS

Landsat ETM+ data were sourced from the United States Geological Survey. Airborne geophysics data were obtained from Mineral Resources Tasmania and soil geochemistry data were collated by Aberfoyle Resources Ltd. This research was conducted at the Australian Research Council Centre of Excellence in Ore Deposits (CODES) under Project No. P3A3A. M. Cracknell was supported through a University of Tasmania Elite Research Ph.D. Scholarship. A. McNeill publishes with the permission of Mineral Resources Tasmania.

REFERENCES

Bierlein, F.P., Fraser, S.J., Brown, W.M., and Lees, T., 2008, Advanced Methodologies for the Analysis of Databases of Mineral Deposits and Major Faults. *Australian Journal of Earth Sciences* **55**, 79-99.

Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M., and Stum, A.K., 2008, Landsat Spectral Data for Digital Soil Mapping. in A. E. Hartemink, McBratney, A. B. and Mendonça-Santos, M. de L. (ed.) *Digital Soil Mapping with Limited Data*, Springer Netherlands, 192-202.

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

Corbett, K.D. and Komyshan, P., 1989, Geology of the Hellyer - Mt Charter Area. Tasmanian Department of Mines, 39.

de Carvalho Carneiro, C., Fraser, S.J., Penteado Croacutesta, A., Moreira Silva, A., and de Mesquita Barros, C.E., 2012, Semiautomated Geologic Mapping Using Self-Organizing Maps and Airborne Geophysics in the Brazilian Amazon. *Geophysics* **77**, K17-K24.

Durning, W.P., Polis, S.R., Frost, E.G., and Kaiser, J.V., 1998, Integrated Use of Remote Sensing and Gis for Mineral Exploration - Final Report. Affiliated Research Center, San Diego State University, 25.

Fraser, S.J. and Dickson, B.L., A New Method for Data Integration and Integrated Data Interpretation: Self-Organising Maps. *5th Decennial International Conference on Mineral Exploration, Expanded Abstracts*, 907-10.

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

Japkowicz, N. and Stephen, S., 2002, The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis* **6**, 429-50.

Kohonen, T., 1982, Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics* **43**, 59-69.

McNeill, A.W., de Bomford, R., and Richardson, S.M., 1998, Relinquishment Report - El 106/87 - Lake Mackintosh. Mineral Resources Tasmania, Internal Report No. Mackintosh 68, 34.

Richardson, S.M., 1994, Exploration Licence 106/87 Lake Mackintosh, Tasmanian Progress Report for the Period April 1993 to February 1994. Mineral Resources Tasmania, 21.

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.

Waters, J.C. and Wallace, D.B., 1992, Volcanolgy and Sedimentology of the Host Sequence to the Hellyer and Que River Volcanic-Hosted Massive Sulfide Deposits, Northwestern Tasmania. *Economic Geology* **87**, 650-66