MINERAL RESOURCES AND DATA61
www.csiro.au

# NWMP Data-Driven Mineral Exploration and Geological Mapping

**David Cole, Lachlan McCalman, Vasek Metelka, Alexander Otto, Jess Robertson, Andrew Rodger, and Daniel Steinberg**

# Summary

The process of exploring for mineral resources consists of several interconnected steps, where the probability of finding a deposit or target is evaluated based on the geological, geochemical and geophysical surveys conducted in an area. All of the acquired data form the foundation of an integrated approach to geological mapping and exploration targeting. The overall aim of the project was to investigate how aspects of machine learning can support the geologist in order to make quicker and more informed interpretations by semi-automated or automated analysis of multi-source datasets.

We have conducted several experiments to assess the utility of different machine learning algorithms applied to freely available data covering the North West Minerals Province (NWMP) in Queensland. The first experiment focused on accessing the copper and gold mineral prospectivity of the Quamby area. The second suite of experiments aimed at the digital geological mapping of the Eastern Succession in the Mount Isa Inlier. Finally, we attempted to predict the concentration of Cu, Au, Pb, Zn indicating possible mineralisation in bedrock as if no cover existed, using *in situ* soil samples.

In prospectivity mapping, we compared the traditional *weights of evidence* (WoE) approach to a classical machine learning method. In general, deriving an accurate prospectivity map based on a small number of known mineral occurrences is very difficult. The result showed, however, that a machine learning approach has fewer restrictions and improved performance when compared to the WoE method, and should be given further consideration.

Several of experiments around the production of geological maps were conducted, including approaches to automated and semi-automated mapping as well as data-driven validation of manually produced maps. The results show that data-driven approaches can extract geological information from remote sensing and geophysical datasets in combination with surface observations. A limiting factor to the performance in deriving a complete geological map was having a sufficient and representative training dataset. Promising results were obtained in the case of focussing on the automated classification of outcropping granites, in particular. In addition to automated mapping, an anomaly

detection algorithm was applied to demonstrate the utility of geophysical data in validating existing geological interpretations but highlighting inconsistencies in the map with respect to available data.

Finally, the challenge of using geophysical datasets, in combination with soil assay samples, to gain insight into the surface and subsurface geochemistry across the region was considered. The supervised machine learning model was able to identify many of the anomalies relating to known areas of mineralisation. This experiment, in particular, highlighted the issue of training data bias, which is common to almost all problems in this space. Here we used a data-driven method to identify areas which are outside where the model has been sufficiently trained. There is also potential to utilise machine learning earlier, to inform the data-collection process, in order to mitigate sample bias and improve model quality and consistency.

While limited to a single region of study, this project has demonstrated the value of information obtained within large, freely available datasets and the potential of machine learning to extract that information in an automated way to improve existing processes. The geochemical element predictions successfully uncovered links between deep geophysical and surface geochemical signatures which can be utilised as an alternative to traditional prospectivity mapping while the anomaly detection and geological classifications provide a more objective assessment for mapping efforts both historical, in production, or planned.

The development and use of new data integration and analysis techniques will not only allow for easier, faster, and more repeatable interpretation of multi-source data but can also improve future data collection practices, all to maximise the value of information gathered for a region of interest.

# Contents

# List of Figures

List of Figures

# List of Tables

# 1. Introduction

The process of mineral exploration consists of several interconnected steps, where the probability of finding a deposit or target is evaluated based on the geological, geochemical and geophysical surveys conducted in an area. In this process, new geological maps and interpretations are generated, including all of the information needed to describe the critical elements of a mineral system [1]. Apart from the derived products, standalone geochemical and geophysical maps that highlight any anomalous regions are usually also employed to provide "unbiased" primary data further characterising the key features.

Geological mapping has since the early days of airborne photography benefited from the continuous cover of remote sensing data and the broad regional context it provided to observations acquired on the ground. Photogeology used the characteristic landform features and colours to discern major rock groups or identify fault lines [2]. Satellite sensors like Landsat and ASTER[1] further enhanced the capability of mapping geology remotely by acquiring data beyond the visible spectral range, which allowed for direct detection of mineral assemblages contained in exposed rocks [3]. Similarly, radar interferometry and photogrammetric processing of stereo images supply digital elevation models that can be utilised in morphometric descriptions of landforms [4]. Most of the surface of the Earth is however covered by vegetation and a variably thick weathering sheet that hinders our geological analyses as we cannot directly observe or sample the underlying rocks and the landforms that we observe cannot be related to the primary geological structures. In Australia, weathering represented by regolith but also younger sedimentary sequences obscure around 80% of the total area. To map the geology underneath this masking veneer, geophysical methods that can image the properties of the subsurface have been routinely used [5, 6, 7]. Most available of the different methods utilised but also providing the best spatial coverage are airborne magnetometry, gamma-ray spectrometry, and surface gravimetry.

[1] Advanced Spaceborne Thermal Emission and Reflection Radiometer.

In magnetometry, we can exploit the magnetisation or susceptibility of various rock types as well as fault and shear zones up to the Currie depth (maximum depth of magnetisation). Gravimetry, on the other hand, maps the differences in densities of lithological units and can theoretically be used to infer the density of the whole Earth mass. Applying filters to data acquired from both of these methods allows us to, for example, separate information coming from different depths or enhance directional gradients indicative of lithological boundaries and faults. Gamma-ray spectrometry has the shallowest depth penetration, of about 50 cm, and measures the abundance of radioelements contained in rocks and regolith material. Both magnetics and gravity effectively map the subsurface and integrate responses from different depths, while gamma-ray spectrometry describes mostly the exposed surface where the primary source of the signal comes from. The gamma-ray spectrometry data may be used directly or enhanced in a comparable way to the multispectral sensor data to derive radioelement concentration maps or ratios.

Integration of the different data has always been a critical process in geological interpretation and derivation of geological maps. Such integration can be done purely based on a knowledge-driven basis with a visual interpretation, or through a numerical overlay operation, through physical modelling (e.g., joint inversion), and finally through a statistical/machine learning process. Data-driven prediction techniques differ from more traditional model-driven techniques primarily in the degree to which they incorporate prior information such as known physical laws into the inference process. At the cost of not having access to exact physical models, they gain the ability to scale efficiently to large data sizes and to quantify their predictive uncertainty. Data-driven approaches can also be applied in problems for which a well-quantified physical model does not exist or is intractable to compute. However, they are extremely limited in cases where there are few or no "ground truth" examples. Much of the recent development in data-driven algorithms has come from machine learning. Machine learning is a field with elements of mathematics, statistics and computer science and the goal of building practical algorithms for learning from data. As such, it has developed powerful tools for regression, classification and density estimation that are useful in problems like the mineral deposit location, geochemical element concentration prediction, or lithological classification [8, 9, 10].

The overall aim of the project was to investigate how aspects of machine learning can support geologists in order to make quicker and more informed interpretations. We have conducted several experiments to assess the utility of different machine learning

algorithms applied to freely available data for the North West Minerals Province in Queensland. The first experiment focused on the comparison of the traditional *weights of evidence* (WoE) method to logistic regression in the Quamby area in assessing copper and gold mineral prospectivity of the Quamby area. In the second experiment, we attempted to predict concentrations of Cu, Au, Pb, and Zn using *in situ* soil element concentrations as training and geophysical data as explanatory variables in order to find anomalous regions suitable for exploration targeting. The third suite of experiments evaluated several approaches to the automated geological mapping of the Eastern Succession in the Mount Isa Inlier. We tried to provide an insight on how we could enhance the abilities of a geologist by a data-driven interpretation as an aid in the mapping process. We attempted to classify lithological units, discover undetected granite outcrops, and find anomalous regions in the interpreted geological maps based on a multivariate dataset.

# 2. Geology of the region

## 2.1. Regional Geology

The Mount Isa Inlier is subdivided into three north-south trending zones, the Western Fold Belt, the Kalkadoon-Leichhardt Belt (KLB), and the Eastern Fold Belt (EFB) [11]. The Eastern Fold Belt has been further divided into geological domains based on geochronology, geophysical properties and structural evolution [12]. The stratigraphy has been traditionally divided into three cover sequences that were deposited on top of the basement represented in the EFB by the Plum Mountain Gneiss which was deformed and metamorphosed by the Barramundi orogeny (~1900-1870Ma, [13]). The deposition of the younger volcano-sedimentary sequences stopped by the early Isan orogeny at ~1600-1570Ma [11]. Subsequent deformation events were accompanied by intensive magmatism, crustal wide hydrothermal fluid flow, and deposition of precious and base metal deposits.

The studied area is mostly situated in the Eastern Fold Belt/Eastern Succession and its later cover sequences but incorporates parts of the Kalkadoon-Leichhardt Belt in its western part (Figure 2.1). The Kalkadoon-Leichhardt Belt constitutes the oldest rocks in the region. In the exposed regions to the west, the belt is composed of the Leichhardt Volcanics which are intruded by the Kalkadoon Granite Suite [14]. The basement rocks of the Eastern Fold Belt, which include at least in the west the lower sequences of the Kalkadoon-Leichhardt Belt, are unconformably overlain by the extension related Magna Lynn Basalt and the Argylla Formation, the oldest lithologies recorded in the Eastern Fold Belt. Slightly older or coeval are in the KLB rocks of the Haslingden Group. These are followed by Myally Subgroup [14]. Overlying the Argylla Formation in the EFB are the Boomarra Metamorphics [15]. In the eastern and southern part of the Eastern Fold Belt, the basement is covered directly by the Bulonga Volcanics that are followed by the mafic Marraba Volcanics, considered part of the Malbon Group. Overlying the volcanic rocks are the sag-phase related Mitakoodi (Malbon Group) and Ballara Quartzites. This

sequence is conformably overlain by the shallow marine Corella Formation, which is in the central region separated from the underlying Mitakoodi Quartzite by the distinctive Overhang Jaspilite [15]. In the KLB, the Quilalar Formation follows the Myally Subgroup and is considered equivalent to the Mitakoodi/Ballara Quartzites and the Corella Formation [14].

After the deposition of the Corella Formation occurred another extensional period in which the Wonga and Burstall Suites intruded into an extensional shear-zone [14]. The Mount Fort Constantine Volcanics probably represent a shallow expression of the Wonga event. The stratigraphic position of the Double Crossing Metamorphics, previously considered basement [16], or co-eval with the Argylla Formation [17, 14] are now considered part of a Wonga-aged core-complex [15].

The renewed sedimentation in the region started by the deposition of the Bigie Formation in the KLB and later on, in the EFB, the Staveley Formation that is at places followed by the cross-bedded Roxmere Quartzite. The Lower Mount Albert Group rocks are considered temporal equivalents to the Staveley-Roxmere sequence [15]. The deepening of the sedimentary environment is indicated by the accumulation of the Soldiers Cap and Kuridala Groups which contain late volcanic members represented by the Toole Creek Volcanics and Hampden Slate, respectively. In the west, the Upper Mount Albert Group rocks now called Mount Roseby Schist [15] are considered coeval with the Soldiers Cap and Kuridala Groups. Further to the west in the Kalkadoon-Leichhardt Belt, the Mount Isa Group rocks were deposited at roughly the same period [14]. Milo Beds found in the central part of the studied region and the Quamby Conglomerate are one of the youngest parts of the Mount Isa Inlier. Probably the last pre-Isan magmatic event was the intrusion of the Tommy Creek Microgranite into the Corella Formation.

After the sedimentation of the Quamby conglomerate, the Isan Orogeny affected most of the rocks in the study area. In the later stages of the orogeny, extensive tonalite-trondhjemite-granodiorite intrusions of the Maramungee Suite were emplaced [11]. The magmatism continued well after the main deformation events and caused wide-spread post-orogenic intrusions consisting of the highly fractionated Williams Supersuite. The younger sedimentary cover consists in the area of the Proterozoic to Paleozoic Georgina Basin as well as the younger Jurassic to Cretaceous Eromanga, Carpentaria and Millungera basins [15].

## 2.2. Mineral exploration and geochemistry

First mining activity in the area was recorded around the 1860s. Protracted exploration operations throughout the past 150 years resulted in the discovery of numerous ore deposit such as Ernest Henry (Cu-Au), Dugald River (Pb-Zn), and Mary Kathleen (U). Alongside the exploration activity, many research institutions and the Geological Survey of Queensland focused on the understanding of the complex geological history of the Mount Isa Inlier. Detailed geological mapping and integration with geophysical datasets were utilised to confirm the lithostratigraphic framework and tectonometamorphic evolution of the region. The exploration activities in the studied area included numerous geochemical surveys which sampled rocks, soils and stream sediments. Except for the stream sediment samples, most of the soil and rocks samples concentrate on discrete exploration targets and are not equally distributed.

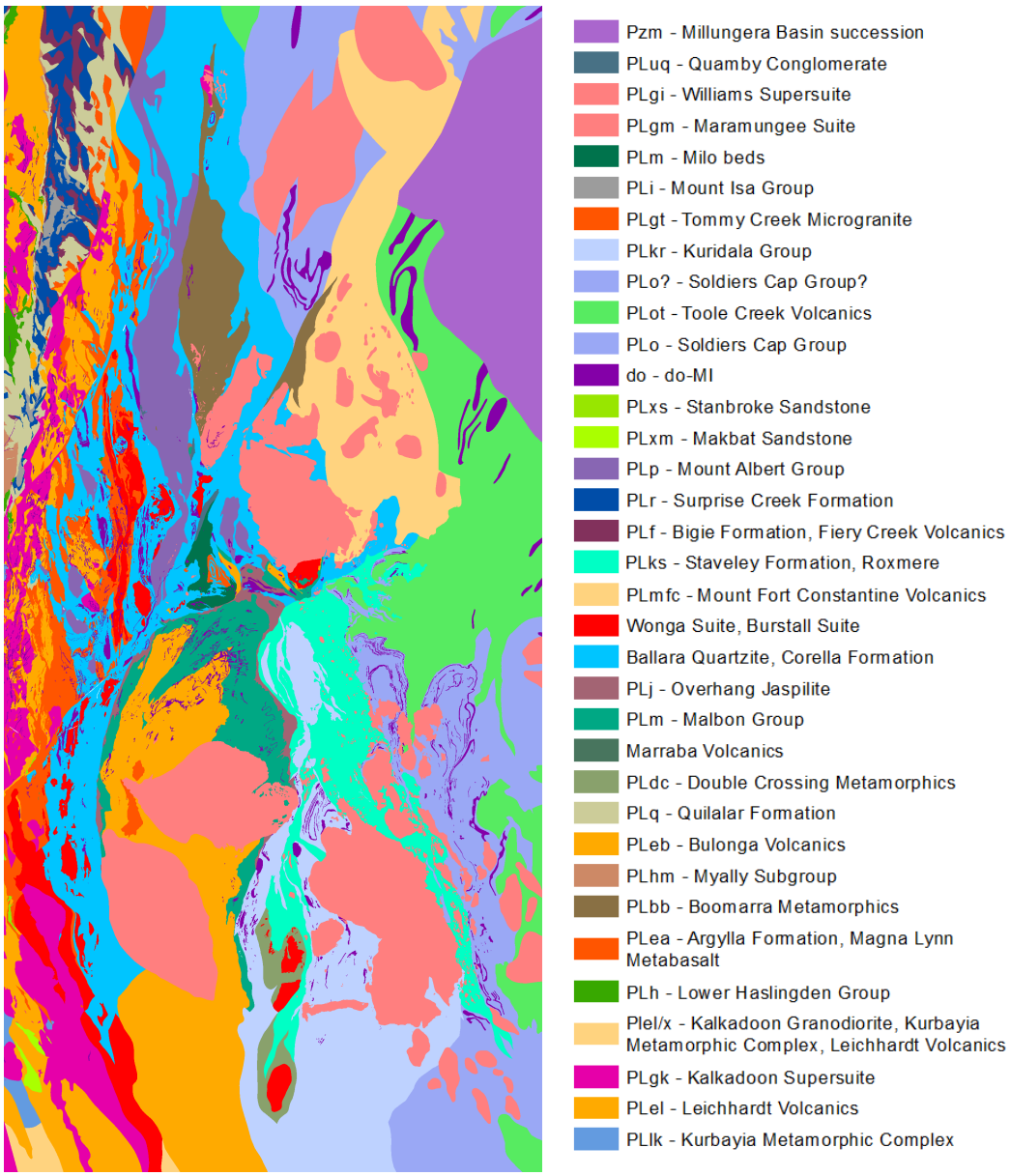



Figure 2.1.: Simplified 1:500,000 solid geology interpreted by the Geological Survey of Queensland (GSQ), unpublished digital data.

# 3. Data-driven modelling

Data-driven prediction techniques differ from more traditional model-driven techniques primarily in the degree to which they incorporate prior information, such as known physical laws, into the inference process. At the cost of not having access to exact physical models, they gain the ability to scale efficiently to large data sizes and to quantify their predictive uncertainty. Data-driven approaches can also be applied in problems for which a well-quantified physical model does not exist or is intractable to compute. However, they are extremely limited in cases where there are few or no "ground truth" examples. For example, an unconstrained gravity inversion would not be possible for a data-driven algorithm as there are no examples of density which a model can learn from.

Much of the recent development in data-driven algorithms has come from machine learning. Machine learning is a field with elements of mathematics, statistics and computer science and the goal of building practical algorithms for learning from data. As such, it has developed powerful tools for regression, classification and density estimation that are useful in problems such as those considered in this report related to understanding the geology in the North West Minerals Province (NWMP).

| Technique | Data-driven | Model Driven |
|---:|---|---|
| uses domain knowledge | + | +++ |
| robust uncertainty | + | + / +++ |
| large predictions | +++ | + |
| large data volumes | +++ | ++ |
| small data volumes | + | +++ |

Table 3.1.: A qualitative comparison of data-driven and model-driven techniques.

## 3.1. Machine Learning

Machine learning prediction algorithms are a subset of data-driven approaches whose objective is to make accurate predictions on *unseen* data [18, page 2]. That is, machine learning algorithms aim not just to fit available data, but to generalise. Algorithms that are designed to provide uncertainties in their predictions attempt to ensure that uncertainty is well-calibrated on unseen data (i.e. that when the algorithm is wrong, it is also uncertain).

The type of algorithms useful for making predictions for the problems discussed here are so-called supervised algorithms: regression, classification and density estimation. These algorithms learn a mapping[1] from some "feature" data $X$ available everywhere in the region of interest to a "target" variable $Y$. In the case of prospectivity analysis, these features $X$ are geophysical and geochemical data interpolated onto a three-dimensional grid, whilst the targets are the presence or absence of a mineral deposit at each grid location. In the case of automated geology classification, $X$ is available geospatial data (geophysics, hyper-spectral data, digital elevation model (DEM), etc.) and $Y$ is a rock type or geological unit.

By default, machine learning algorithms do not explicitly encode a physical model between observations and the unknown state of the world. Such domain knowledge can instead be encoded by "feature engineering". This process involves applying transformations to input features $X$ which encode domain knowledge or expertise. Examples of this are upward continuations of gravity data and gradient operations on magnetics.

## 3.2. Evaluating Data-driven Models

To test the generalisation of a model (that is, how well it predicts data it has not seen) the standard approach in machine learning is to test accuracy on data withheld during model training.

When dealing with small amounts of data, a cross-validation approach removes only a small fraction of the data to evaluate model accuracy, then repeats the training and validation with different small fractions removed and averages the results.[2]

[1] Depending on the approach and algorithm choice, the mapping can range from a simple linear function to a complex non-linear mapping defined by thousands of learnt parameters.

[2] The extreme case of this is "leave-one-out" cross-validation where a single data point is used to evaluate

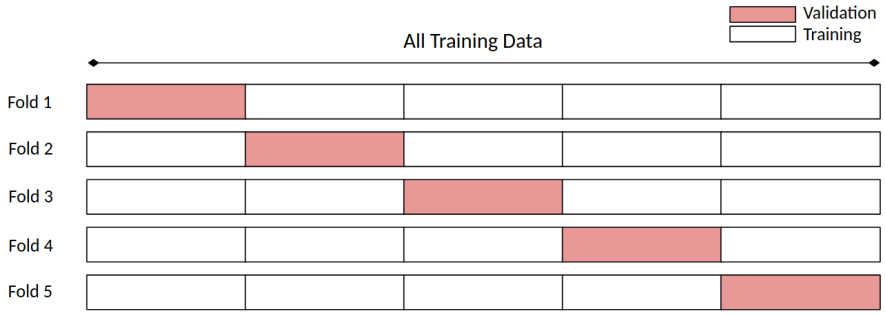



Figure 3.1.: An illustration of how to split the data into 5 folds for cross-validation. Usually, the data is randomly sorted before splitting into these folds.

A common mistake in evaluating the quality of predictive models is to look at how well they predict the data provided to them for learning. This metric favours a model that memorises the training data but can not generalise, a problem referred to as *over-fitting* [18, Chapter 1.3].

## 3.3. Performance Measures

There are many measures we can use in a cross-validation framework to asses the performance of an algorithm. These measures mainly fall into two categories: (1) those that measure the degree to which the predicted outcome matches the value of the held-out target value, and (2) those that measure how well the algorithm estimates its confidence in its predictions.

For classification, standard measures of the predicted outcome are accuracy, precision and recall. Accuracy is a measure of the rate at which the classifier correctly predicts the class label. Precision is a ratio of how many predicted items are true positives (true positives / all predicted positives). Recall is a ratio of how many of the true positives have been detected (true positives / all positives in the testing data). Another measure, $F_1$ score, can be used as a measure of both precision and recall in a single value.

To measure the outcome in regression, we typically use mean-squared error or the *coefficient of determination* ($R^2$), which is related to standardised error.

the model and the training process is repeated for all data points.

Unfortunately, it can be hard to ascertain the algorithm's confidence in its predictions by using only the aforementioned measures. To obtain a measure of the algorithm's confidence in its predicted outcome, we need to be using a probabilistic algorithm that returns a probability density for its predictions (from which we can, for example, obtain confidence intervals). With this, we can evaluate the probability of the *held-out* testing data under this distribution. Utilising this measure, we look for models that can predict the held-out data *with high probability*, in Figure 3.2 we demonstrate different scenarios in which different levels of confidence and correctness can influence this measure.

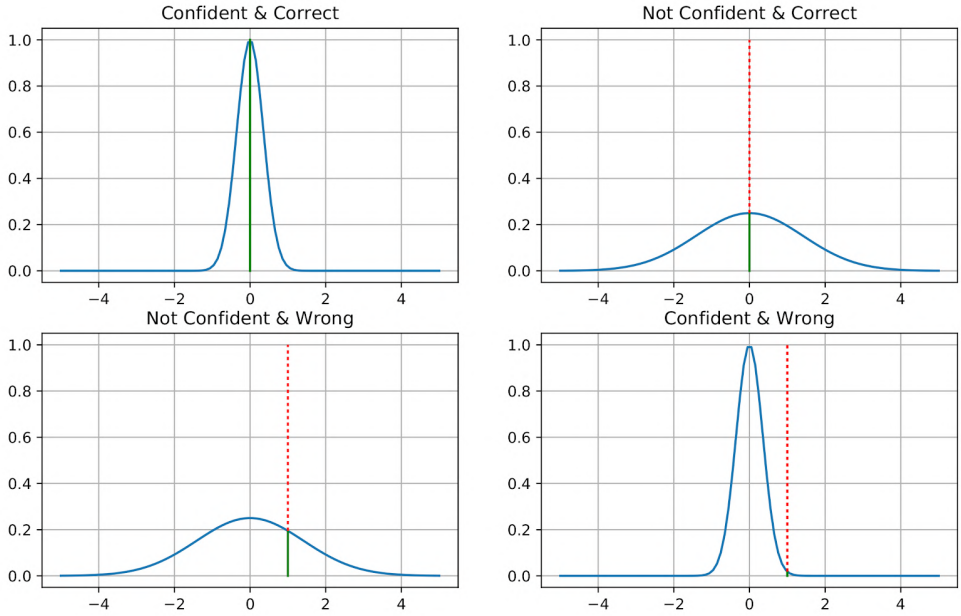



Figure 3.2.: A demonstration of how we can use the predictive probability density from a probabilistic regressor to measure its confidence. Ideally, we would like a confident and correct prediction, and we would prefer an unconfident and correct or incorrect result to a confident and incorrect result.

# 4. Datasets and processing

In both the geochemical predictions and the geology classification experiments, a number of datasets were extracted and processed. All of the utilised data is considered pre-competitive and freely available to explorers. Prior to using geospatial data to train a machine learning model, it must be processed to account for projections, extent, resolution, missing values, etc. Additional features were also derived from the data, such as wavelet transformations, textural features, and distance-to-line features. This section describes all of the utilised data along with pre-processing steps applied to it.

The extended area of interest (Eastern Succession of the Mount Isa Inlier) was defined with the boundary shown in Table 4.1. The bounds were chosen based on similarity and continuity of the lithological units. All data sets were transformed into a common coordinate reference system (CRS) and aligned on a grid with resolution chosen by the highest resolution dataset.

| | |
|---|---|
| **Coordinate reference system** | `+proj=longlat +datum=WGS84 +no_defs` |
| **Bounds** | [139.696, -22.220, 141.085, -19.207] |
| **(Quamby bounds)** | [140.099, -20.733, 141.000, -20.000] |
| **Image size** | $3616 \times 1667$ |
| **Resolution** | 0.00083325 deg/pixel |

Table 4.1.: All input/output data used for modelling was transformed (and rasterised) to the following image specifications.

## 4.1. Datasets

### 4.1.1. Geophysics

All of the geophysical data were acquired from the QDEX Data system of the Queensland Geological Survey. As the predictions and classification access the regional geological and geochemical features, regional-scale data were used, providing appropriate resolution and level of pre-processing.

### 4.1.2. Gravity

The Bouguer anomaly gravity grid represents a gridded compilation of all freely available gravity observations from exploration companies and state and federal regional surveys maintained by the GSQ [19]. A value 2.67 $g/cm^3$ was used to produce the Bouguer corrected values. The gridding at 1 km spatial resolution used a standard minimum curvature algorithm.

### 4.1.3. Magnetics

The reduction to the pole (RTP) total magnetic intensity (TMI) grid, used in all of the experiments, was produced by applying the RTP correction to a merged TMI grid. The TMI grid was created by merging grids created from surveys collected by state and federal governments and companies [20]. The grids were levelled by employing the Australia Wide Airborne Geophysical Survey. The individual grids were gridded using the standard minimum curvature algorithm. The final spatial resolution of the merged grid was set to 90 m. Apart from the TMI the first vertical derivative of the field was also utilised. The gradient filtering was applied to the merged TMI [21]. A one-cell upward continued magnetic grid was also generated to suppress the near surface responses in mostly outcropping regions.

### 4.1.4. Gamma-ray spectrometry

The gamma-ray spectrometry data, utilised as another input, stem from the Geoscience Australia's Third Edition Radiometric Map of Australia, in which grids were merged from individual surveys collected by state and federal governments as well as companies.

The individual grids were created utilising the standard minimum curvature algorithm and levelled using the Australia Wide Airborne Geophysical Survey [20]. The original Total Count, Potassium, Thorium and Uranium concentration grids served as input in the experiments without any further enhancements.

### 4.1.5. Remote Sensing

All of the remote sensing data were sourced internally from CSIRO, but slightly modified versions can be obtained as freely available data from Geoscience Australia and the QDEX Data system of the Queensland Geological Survey.

### 4.1.6. ASTER

The ASTER data utilised in our experiments stem from the 3D Mineral Mapping of Queensland - Version 2 ASTER and related geoscience products [22]. Apart from the Version 2 mineral maps, the standard mineral map products defined by [22] were calculated from the updated reflectance and emissivity mosaic data (Table 4.2) with only the composite mask applied to create spatially continuous data without a large amount of missing values. The Version 2 calculated green and dry vegetation mask were used as separate input data. The raw reflectance and emissivity data were also utilised. The resolution of the ASTER data and mineral maps varied between 30 and 90 m, corresponding to the short-wavelength infrared (SWIR) and thermal infrared (TIR) sensor spatial resolution.

### 4.1.7. SRTM digital elevation model

The digital elevation model data were obtained from the 1 arc second DEM of bare-earth with attempted removal of vegetation and man-made structure effects [23]. Apart from the DEM, the multi-resolution valley bottom flatness (MrVBF) grid was used to assess the general erosional or depositional setting of the area. The MrVBF was produced from the smoothed version of the DEM using the algorithm of [24].

### 4.1.8. Geology and Geochemistry

The spatial geological and geochemical data were acquired through the Geological Survey of Queensland and can be downloaded from QDEX Data or QSpatial. Shapefiles that were used include geological maps from the package Queensland geological mapping digital data – November 2012 [25]. A simplified version of the 1:500,000 scale solid geological map was obtained directly from GSQ(unpublished). This simplified version was used in most of the classification experiments. Apart from the geological maps, GSQ Site Observation data [26] containing point locations of lithologies observed in the field were used directly as training points.

Geochemical data were sourced from the Mount Isa East 2016 Data Package [27] which contains information extracted from open-file company reports. Only the soil sample analyses were used in our experiments. The data were filtered to only include samples from one fraction size, less than 180 um. Below detection limit values were replaced by half of the detection limit.

## 4.2. Raw Data Files

Below is a list of the raw geospatial datasets available for the region used for modelling.

| Data | Files |
|---|---|
| ASTER | emissivity90mV2.tif<br>reflectance30mV2.tif |
| dem | dem1sv1_0.tif |
| gravity | qld_gravity_grid.tif |
| magnetics | Qld_merged_mag_1VD.tif<br>rqld_VRTP_v6.tif |

| | |
|---|---|
| mineral maps | AlOH_group_compositionV2_nomask.tif<br>AlOH_group_contentV2_nomask.tif<br>FeOH_group_contentV2_nomask.tif<br>Ferric_oxide_compositionV2_nomask.tif<br>Ferric_oxide_contentV2_nomask.tif<br>Ferrous_iron_indexV2_nomask.tif<br>Gypsum_indexV2.tif<br>Kaolin_group_indexV2_nomask.tif<br>MgOH_group_contentV2_nomask.tif<br>Opaque_indexV2_nomask.tif<br>Quartz_indexV2.tif<br>Silica_indexV2.tif |
| mineral maps masked veg corrected | AlOH_group_compositionV2.tif<br>AlOH_group_contentV2.tif<br>c2230_indexV2.tif<br>FeOH_group_contentV2.tif<br>Ferric_oxide_compositionV2.tif<br>Ferric_oxide_contentV2.tif<br>Ferrous_iron_indexV2.tif<br>Kaolin_group_indexV2.tif<br>MgOH_group_contentV2.tif<br>Opaque_indexV2.tif |
| mrvbf | mrvbf_int.tif |
| gamma-ray spectrometry | k_pct_filtered.tif<br>TC_filtered.tif<br>Th_ppm_filtered.tif<br>U_ppm_filtered.tif |
| solid geology | NWQ_simplified_solid_geology_polygons.shp |
| vegetation | Green_vegetation_contentV2.tif<br>Composite_maskV2.tif |

## 4.3. Feature engineering

In addition to the raw data (cropped and transformed) described above, a number of derived feature datasets were generated to extract additional information from the original data sources. These datasets were included as inputs into the models described in subsequent sections of this report.

### Magnetic line data

Linear features were automatically extracted from the variable reduction to the pole (VRTP) magnetic intensity image using the CET Grid Analysis extension for Oasis Montaj. The lines were extracted separately for four directions striking N, NE, E, SE to enable separate evaluation of faults with possibly different kinematics connected to various deformation events. After extracting linear features from the magnetic data, a *distance-to-line* value was derived across the area to provide an additional scalar covariate values to be used as an input to data-driven models. These four distance measures provide spatial context to gradients in the magnetic data, mapped faults and lithological boundaries.

### Wavelets transformation

Wavelet transformations of geophysical data (magnetics and gravity) are particularly useful in that they relate depth information of the geological features which produced the signal. Wavelet decompositions were performed on the magnetics and gravity data across a range of eight scales and four directions[1] to produce 32 additional covariate features per geophysical input.

### Textural features

The textural properties of geospatial datasets were used to inform a manual assessment of the data to generate interpreted products (e.g., textural properties in magnetics to inform geology interpretation). To automate this process, Grey-level Co-occurrence

[1] Wavelet decompositions were evaluated using python's `pyfftw` fast Fourier transform library. The 8 scale parameters were: 0.006, 0.012, 0.026, 0.054, 0.113, 0.235, 0.490, and 1.022. The 4 directions were 0, $\pi/4$, $\pi/2$, and $3\pi/4$.

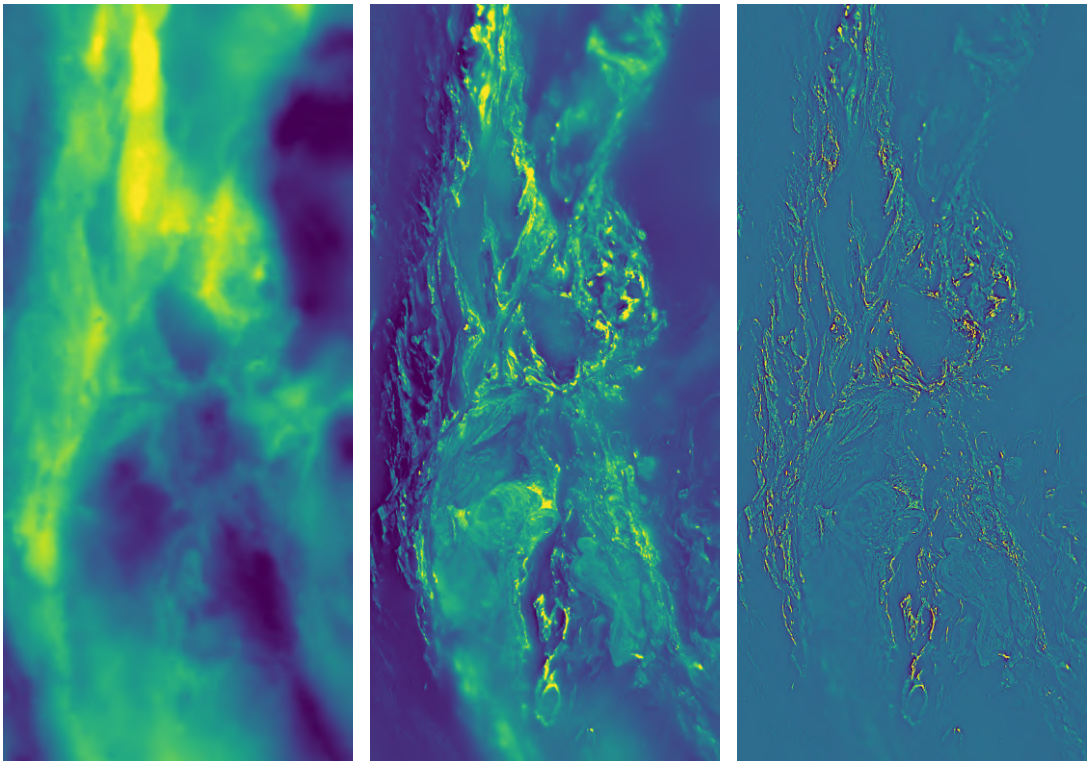
Matrix (GLCM) textural properties were derived to encode the local texture as a covariate used as a model input (See Appendix A for details).

### DEM features

In addition to the MrVBF [24] data, which is available for all of Australia, the DEM was used to derive both Terrain Ruggedness Index (TRI) and Topographic Position Index (TPI) features in order to capture secondary properties of the terrain which may correlate with the geology of the region in outcropping areas. TRI and TPI are measures of the relative elevation of a point with respect to neighbouring pixels [28].

## 4.4. Modelling under cover

While a large proportion of the modelled region is outcropping, most of the area remains under cover. The aim of some of the problems addressed in this report was to make inferences on geological structures or properties present in those areas under cover based on surface observations. For these models, geospatial covariates were limited to geophysical datasets which measure properties below the surface. For the NWMP area, we used three such datasets: gravity, magnetics VRTP, and magnetics first vertical derivative (1VD) (Figure 4.1). Derived features mentioned above, such as wavelets, textures, and magnetic lines were also used.



(a) Gravity. (b) Magnetics VRTP. (c) Magnetics 1VD.

Figure 4.1.: Gravity and magnetic geospatial features provide critical information below cover.

# 5. Prospectivity mapping

One data-driven prediction method already in use by GSQ is that of *weights of evidence* (WoE) used for predicting mineral deposit locations, given geophysical and geological data available for a given area. GSQ have applied WoE in the Quamby region analysis. In this section, we compare WoE to a traditional machine learning approach both qualitatively in terms of how the algorithms work and then quantitatively as applied to the Quamby area prospectivity prediction problem.

## 5.1. Weights of evidence and machine learning

*Weights of evidence* is an empirical statistical technique designed to understand the correlation of features to a binary target variable (presence of mineral deposit in the case of the Quamby analysis). It differs from typical machine learning algorithms in that it is not designed to predict unseen data but to try and understand the relative importance of different features for the presence or absence of the target variable. It is limited to considering presence/absence correlations only, and relaxing this assumption to include the possibility of more complex relationships creates the issue that wildly different feature importances can have the same explanatory power. Such observational relationships should therefore never be construed as causal.

It is possible to use WoE as a predictive tool by computing the probability of a positive label from the importance weights computed for each feature. This probability is a function of the presence or absence of different binary features, and the total number of positive and negative examples in the model. The Quamby analysis carried out this process and produced a probability density estimate of mineral deposits.

As a predictive tool however, WoE has several key limitations compared to other approaches:

- The features must be binary. This requires deciding on a cut-off value and limits the complexity of any relationships to presence or absence.

- The features must be conditionally independent. This is difficult to satisfy when interpretations and derivative products are a source of additional explanatory variables.
- The relationships are independent. There is no ability to consider co-occurrence of features or any more complex modelling.
- The probability estimates are unlikely to be calibrated. The probabilities would be unlikely to generalise to model data that was not given to the algorithm.

Fortunately, there are several machine learning approaches to this modelling problem that overcome all these limitations. The prospectivity modelling problem can be approached with a class of machine learning algorithms called probabilistic classifiers [18, Chapter 4]. These are algorithms which, given a set of labelled (binary or categorical) feature-target pairs, predict the probability of a label at a location based on the features at this location. Examples of these algorithms include logistic regression, Gaussian process classification, and neural network models. These algorithms have a number of advantages over WoE for making predictions.

- Features can be binary, categorical or continuous.
- Features need not be conditionally independent, so derived data products can be added.
- In the case of the Gaussian process and the neural network, complex, non-linear relationships involving multiple features can be learned automatically from the data.
- The algorithms are designed to generalise, that is, to predict unseen data and know the confidence of their predictions.
- In the case of the neural network models, missing data can be probabilistically imputed as part of the training step.
- Training the model is a separate step from prediction so that the underlying relationship is not dependent on the particular voxelisation.

We recommend that, in this and equivalently formulated problems, GSQ use these types of approaches in the future.

It is important to note that the problem as formulated is somewhat ill-posed: we have no negative examples of mineral deposits, nor do we know the sampling distribution used to find the positive examples. To overcome this issue, one can generate a proportion of random points from the region that could contain mineral deposits or not, then interpret the classifier prediction as a density ratio between the "background" distribution and the distribution of known mineral deposits. See Figure 5.1 for a 1-dimensional example. This reasoning could also be used to motivate the WoE approach, and there is no particular need in WoE to use all voxels in the prediction region as negative examples. Controlling the prior density of mineral deposits can either be done by changing the number of random examples (as in the WoE approach) or explicitly setting the desired value in the model.

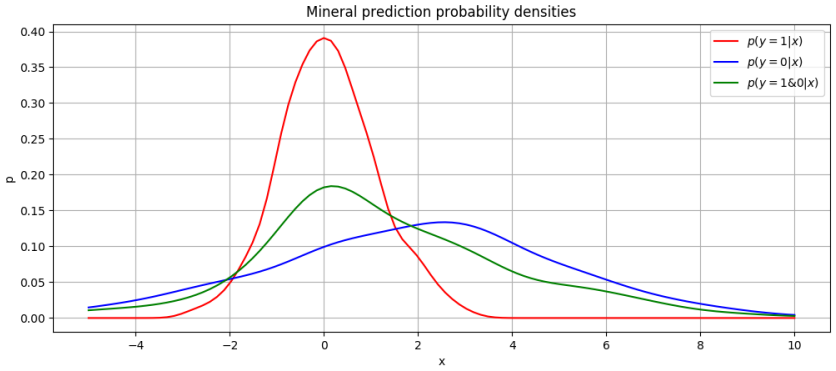



Figure 5.1.: In a regular, one-dimensional, classification task, we wish to learn a classifier that optimally discriminates between the red and blue classes (mineralisation and no mineralisation). However, in this instance we have no examples of the blue class, so we randomly sample from the joint distribution of classes (green), and instead learn a classifier that can distinguish this "background" distribution from the red (mineralisation) distribution.

## 5.2. Quantitative comparison in the Quamby region

A WoE model was developed by GSQ to predict the probability of mineral occurrence across two distinct domains (Constantine and Mary Kathleen) within the Quamby region. Separate WoE models were derived for each domain using locations of known mineral deposits and a 3D model of geospatial properties.

As stated in Section 5.1, WoE modelling has several limitations, including the assumption of feature independence and the reuse of training data in applying the model, which limit its usability as a predictive model.

This section presents an alternative method for generating a probabilistic classifier from the given dataset using a logistic regressor. Given the nature of the specific problem and the data available, the usefulness of this model is also limited. As an illustration, however, the principles described in Section 3.2 are applied here to validate and measure the performance on the model and to compare a machine learning approach to this problem with the WoE method.

### 5.2.1. Model training

As with the WoE model, machine learning approaches require a training dataset of geospatial covariates (or features) $X$ and targets $Y$ (presence/absence of mineralisation). For the Quamby area, the training data consists of the features which are present in the Common Earth Model (CEM) and the targets which are the locations of known mineral deposits.

There are, however, several potential issues within this training dataset which make learning a useful classifier very difficult:

1. **A very small number of *positive* target examples to train on**

   With respect to the number of covariates (8) the number of positive training points is very small (~15 points per domain). This issue tends to lead to a classifier which overfits to the few training points provided and does not generalise well. It also makes testing/validation difficult as the already small dataset needs to be further split into training and testing datasets to avoid evaluating the model on data it has already seen.

2. **Potential for missing positive examples**

   The location of each known mineral deposit/occurrence is represented as a single point/cell within the Common Earth Model. However, large deposits are likely to span more than one cell (250m x 250m x 50m), and it is possible many cells adjacent to the positive sample cells should also be labelled as positive for mineral occurrence. In this case, the training data does not accurately represent the relationship we are trying to model.

3. **No data collected on explicit *negative* examples**

   All cells other than the few positive examples are assumed to be negative (no mineral occurrence) or from a background distribution of unlabelled cells. Many of these cells will have been explored and will be known to not contain the minerals of interest. Other cells will not yet have been explored and will have an unknown status. This information would be very useful to a classifier trained to discriminate between potentially mineralised and no-mineralised areas, which in general requires training data with both positive and negative examples confidently labelled.

4. **Extremely unbalanced dataset**

   The number of negative/other training points (~3 million) is many orders of magnitude higher than the number of positive training points (~15). This issue can lead to a classifier which is biased towards the negative samples. For example, a classifier which predicts 0 for all cells would have excellent accuracy.

Despite the above issues, a logistic regression classifier was trained for each of the two domains, and the results are presented below. An advantage that a logistic regression approach has over WoE is that it does not require the input features to be conditionally independent. Therefore this model was trained using all properties available within the CEM.

In order to evaluate the performance of a classifier, it is necessary to test it on data not seen during training. Given the small number of positive samples, available a "leave-one-out" cross-validation approach with respect, to the positive samples was used. For example, with 14 positive samples, the dataset was divided randomly into 14 folds of equal size, each containing one positive sample. The model was trained on 13 of the folds and then evaluated on the held-out fold. This process was repeated 13

times, with a different fold held-out each time. The results presented below are the aggregated predictions of the held-out data.

### 5.2.2. Model results and performance

The estimated log-probability of the logistic regression approach is shown for the Constantine domain in Figure 5.2a as a series of slices at different depths. Red dots show the location of the 15 training cells. The actual probability estimates made by the model are very small since the implicit prior probability of mineral occurrence is $4 \times 10^{-6}$ (14 positive training cells / 3,466,225 total cells within the domain). By way of comparison, the log-probability of GSQ's WoE (complete) model is shown in Figure 5.2b.

The performance of the classifier in discriminating between positive and negative targets is illustrated by the receiver operator characteristic (ROC) curve [29, Chapter 5.6] (see Figure 5.3). In addition to the GSQ WoE model (`woe_complete`) and the logistic regressor trained using cross-validation folds as described above (`logistic_regressor_xall`), also shown is a logistic regressor trained and evaluated on the entire dataset (`logistic_regressor_xall_no_cv`). As described above, this is bad practise and is misleading in regards to how well the model generalises, yet is it comparable to the WoE mode,l which uses the entire dataset to determine the model weights. As can be seen from the figure and associated area under curve (AUC) values, the logistic regressors perform better as classifiers given the CEM data and known mineral occurrence locations.

Table 5.1.: Comparison of negative log-probability of predictions made by each classifier for the Constantine domain.

| **Classifier** | **log-loss** |
|---|---|
| `logistic_regressor` | 0.0000429 |
| `logistic_regressor_no_cv` | 0.0000390 |
| `woe_complete` | 0.000337 |

Another way to measure the performance of a probabilistic classifier is to look at how well the predicted probabilities match with the true values. For a probabilistic

classifier, we can use the *log-loss* (or *cross-entropy error*) [18, Chapter 4.3], which is the negative log-likelihood of the test sample given the predicted probability (i.e. a measure of how well the predicted probability matches the true label — incorrect predictions are penalised more if they are confident). By definition, this is a loss function, so smaller values indicate better performance. The calculated log-loss of the three classifiers for the Constantine domain are provided in Table 5.1. The logistic regressor performs better than the WoE model, but again, the performance measure is aggregated over all data points in the test set and is, therefore, most heavily influenced by the class with the highest membership (i.e. the negative examples in the unbalanced dataset).

Table 5.2.: Comparison of negative log-probability of predictions made by each classifier for the Mary Kathleen domain.

| **Classifier** | **log-loss** |
|---|---|
| `logistic_regressor` | 0.0000987 |
| `logistic_regressor_no_cv` | 0.0000916 |
| `woe_complete` | 0.000165 |

The same set of classifiers were also trained on the Mary Kathleen domain dataset and the same comparisons between them were made. The predictions of the logistic regression classifier and the WoE classifier at various depths can be seen in Figures 5.4a and 5.4b. The ROC curves are plotted in Figure 5.5 and the corresponding "log-loss" performance measures show in Table 5.2. The relative performance of the logistic regression compared to the WoE is same as for the Constantine domain.

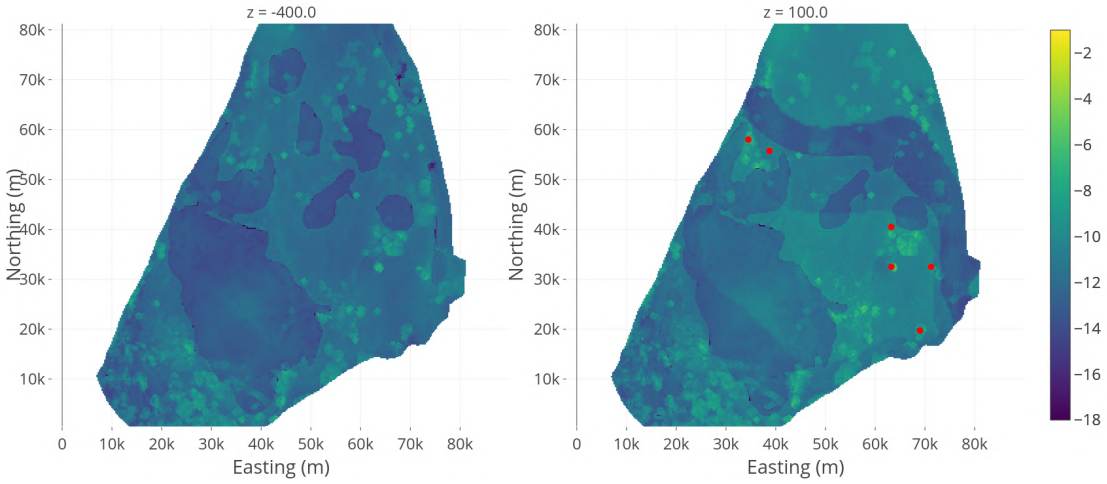
## 5.3. Summary

Traditional machine learning classification methods have a number of advantages over WoE, including that they are designed to generalise to make predictions in areas which they have not been trained on and they do not require conditional independence of features. A logistic regression classifier is one such machine learning model that was applied to this dataset in comparison to WoE. The results show improved classifier

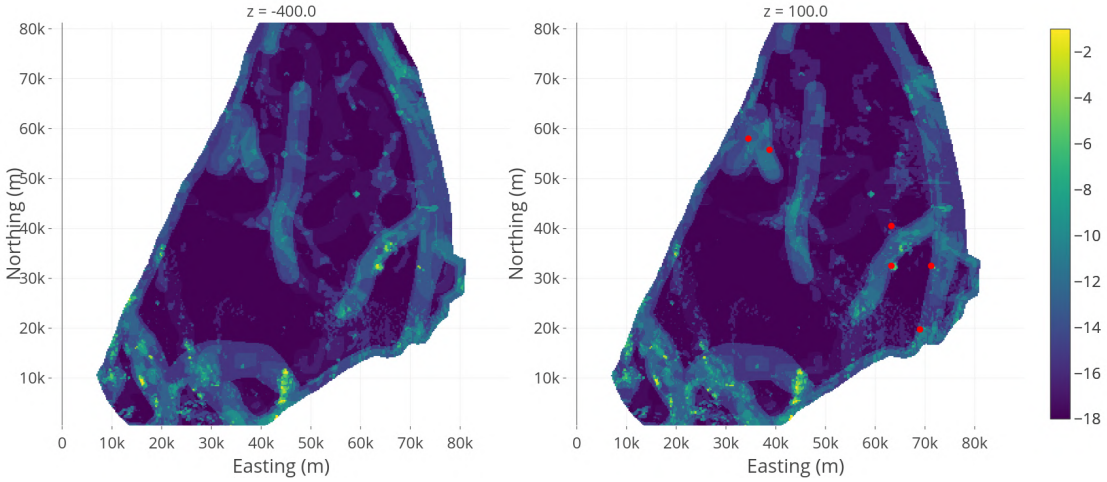
performance as measured by the area under ROC curve and the log-loss of probabilistic predictions.

The particular dataset available for making mineral occurrence predictions within the Quamby region has several of limitations, which means that any derived model will also have limited value. In particular, the very low number of positive samples, potentially missing positive examples, and the absence of explicitly defined negative examples make training a classifier difficult.

As an illustration, it has been shown that the use of machine learning classification has the potential for similar types of problems as an alternative to WoE, mainly if some of the training data issues can be addressed.




(a) Logistic regression model (all CEM features)




(b) GSQ WoE (complete) model

Figure 5.2.: Logistic regression classifier — Log-probability of mineral occurrence for the Constantine domain at two depth slices. Red dots indicate the location of positive training cells.

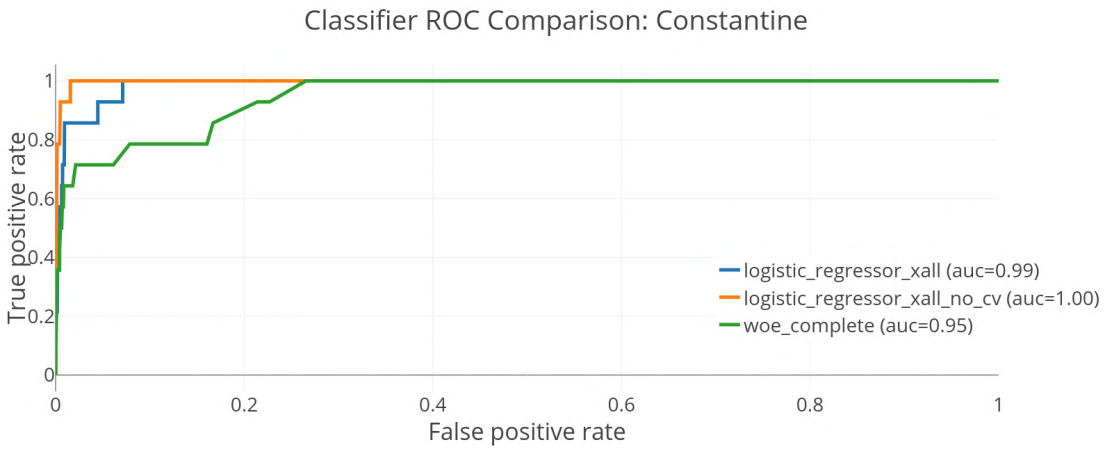
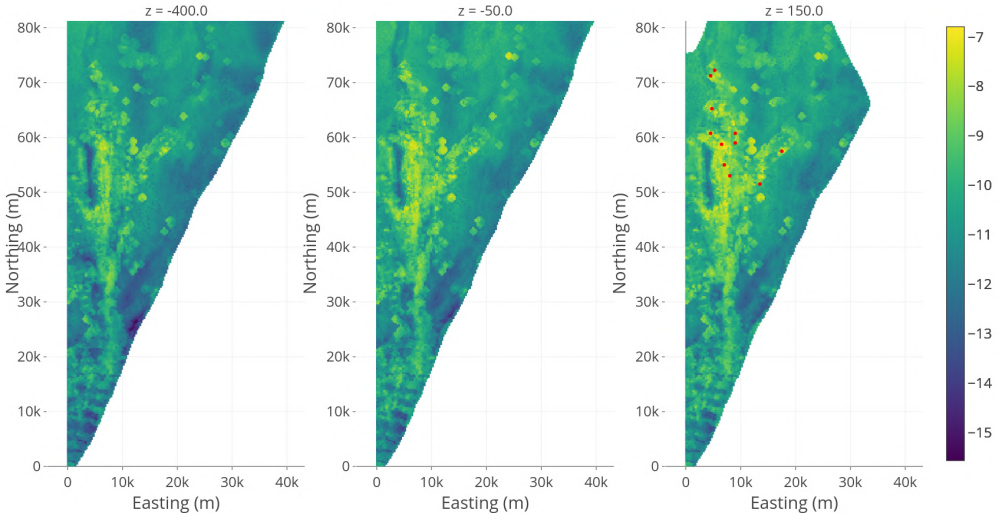
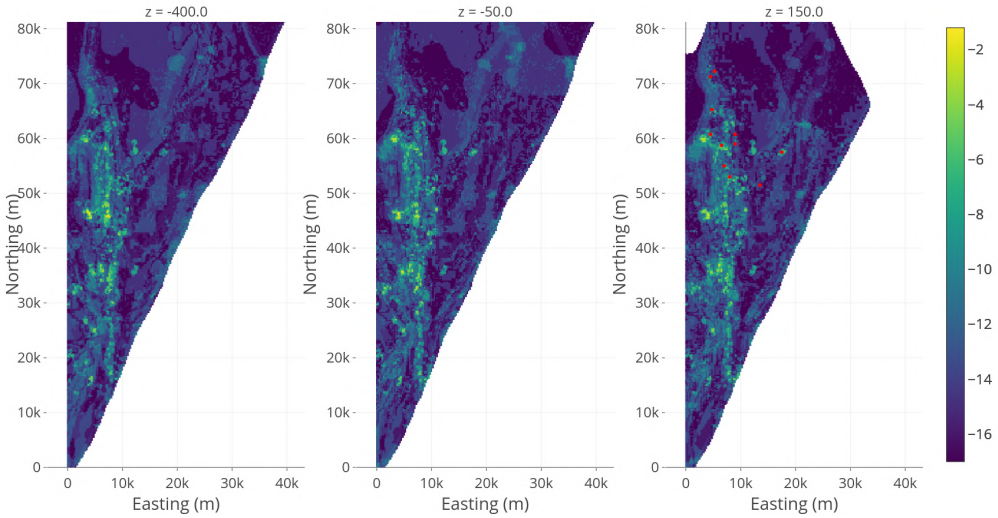



Figure 5.3.: Receiver Operator Characteristics as a performance measure of each classifier for the Constantine domain.




(a) Logistic regression model (all CEM features)




(b) GSQ WoE (complete) model

Figure 5.4.: Logistic regression classifier — Log-probability of mineral occurrence for the Mary Kathleen domain at three depth slices. Red dots indicate the location of positive training cells.

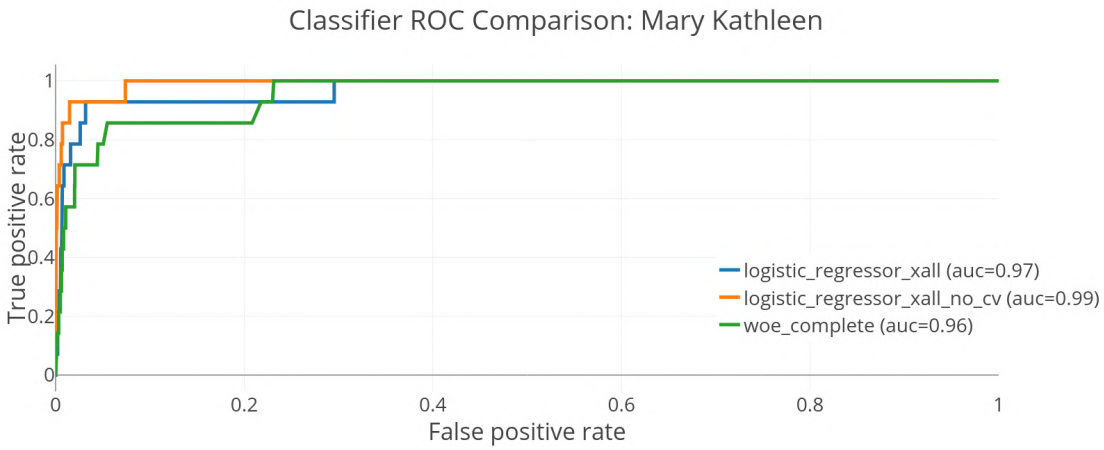



Figure 5.5.: Receiver Operator Characteristics as a performance measure of each classifier for the Mary Kathleen domain.

# 6. Geology classification

We considered two approaches to automating the process of geological classification for the region. The first one (Section 6.1) utilises the existing geological interpretations (a geological map) by drawing samples from it and demonstrating how such information may be used to infer the geology across the wider area. The second approach (Section 6.2) limits the inputs to only the original surface observations made in the region and is hence less dependent on expert interpretations made in the production of the geological map.

## 6.1. Interpreted geology-based classification

The task was to use freely available public data held by the Geological Survey of Queensland (GSQ) to attempt a classification of lithostratigraphic units derived from the simplified solid geology of the Eastern Succession, considered as the target in this case. The aim is to evaluate the predictive possibilities of a suite of classification experiments.

These include the ability of data-driven model to replicate the published geology when: all the possible classes are known *a priori*; all classes are known, but the model is trained on a sub-sample of inputs from each potential class, and lastly when the assumption that the number of actual classes is unknown but a suitable sampling regime is possible.

### 6.1.1. Model

For all of the scenarios investigated, a random forest classifier (RFC) was used [29, 30]. A random forest is a non-linear classification (and regression) algorithm that uses an ensemble of decision trees that vote on the correct prediction to produce an output. Decision trees themselves are a simple prediction algorithm that bisects the training

data along single feature dimensions of maximum variance. By themselves they tend to over-fit, so a random forest creates an ensemble of these decision trees, each of which is only provided with a random subset of the total feature dimensions. For a categorical prediction, the random forest selects the mode of the prediction ensemble. Model parameters used here were 50 estimators in the ensemble and the *class_weight* option set to "balanced" which weights samples to account for the number of samples per class.

The reasons for the selection of this model are the simplicity of implementation and the fact that a decision tree is intuitively understandable to most. Other benefits of this model choice are that the inputs/covariates do not require normalisation and the models return a measure of the feature importance[1] so we might make informed decisions about what input features/covariates provide the most value.

## Covariates

A number of covariates, or inputs, were made available for testing and training. These are broken into the following categories,

1. Gravity
2. Magnetics (RTP)
3. DEM
4. Radiometric (gamma-ray spectrometry)
    - Potassium (K)
    - Thorium (Th)
    - Uranium (U)
    - Total count (TC)
5. ASTER
    - ALOH Composition and Content
    - FeOH Content
    - Fe Oxide Composition and Content
    - Ferrous Iron
    - Gypsum Index
    - Kaolinite Group Index

[1] For Random Forests the *feature importance* is a measure of how much information is gained, on average across all decision trees, by the addition of that feature. Implementations of Random Forests (e.g. `scikit-learn`) provide this measure for a trained model.

- MgOH Grp Index
- Opaque Index
- Quartz Index
- Silica Index

In all cases, the covariates used were in a `tif` file format with spatial dimensions of approximately 3600 by 1600 pixels covering the general Mt Isa Inlier region.

## 6.1.2. Experiments

### Sampling from each class across the entire spatial extent

The RFC was trained by firstly varying the number of (randomly selected) data points used from each class and secondly by restricting the number of covariates used in the model. In this scenario, the total number of classes within the outcropping parts of the solid geology map is fixed and known. Additionally, since each class is sampled by a percentage of the total number of raster pixels within a given class, an unbalanced dataset is created for the training of a given scenario, but as noted previously, this is mitigated by applying a class weighting so that a balanced dataset is ultimately modelled.

The covariate data used in the first experiment are all of those listed above, while the second did not use any ASTER covariates. The percentage of data sampled from each class was 100%, 90%, 50%, 10%, 5% and 1% respectively. Referring to Figure 6.1, we find that, as the number of training points is reduced, the accuracy of the model, as would be expected, is also reduced.

However, the exclusion of the ASTER covariates only leads to an approximate 3% decrease in the overall accuracy of the model and suggests that these additional input layers are not needed. Figures 6.2, 6.3 and 6.4 summarise the results of the experiments showing the importance of covariate layers, comparing the classification with the current geological map and specifying the probability/uncertainty of the predicted classes.

Examining the model-predicted class results (Figures 6.2 and 6.3) when only 1% of the available data from each class is used shows that a 70% accuracy is still attainable even when the inputs are severely restricted. In this case, it suggests that if a field operator was able to identify all potential classes within a given spatial domain successfully, and collect a reasonable amount of ground truth observations, then a RFC model could

be trained that would greatly assist in helping to define the class boundaries of the different geological units within the spatial area of interest.

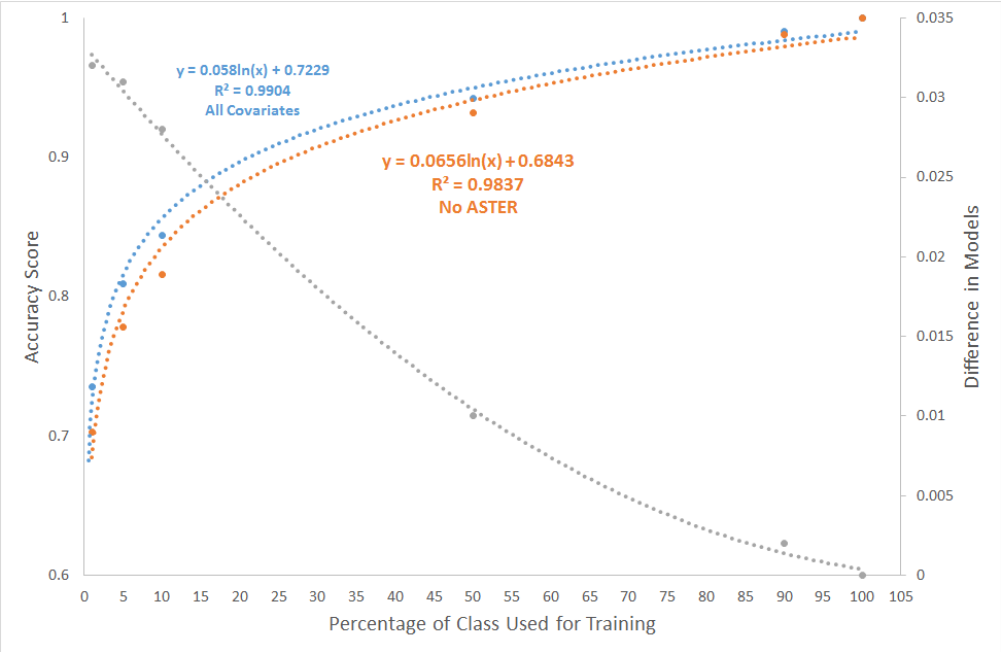



Figure 6.1.: The differences in model accuracy when the number of data points from each class is restricted, and when the number of covariates is restricted.

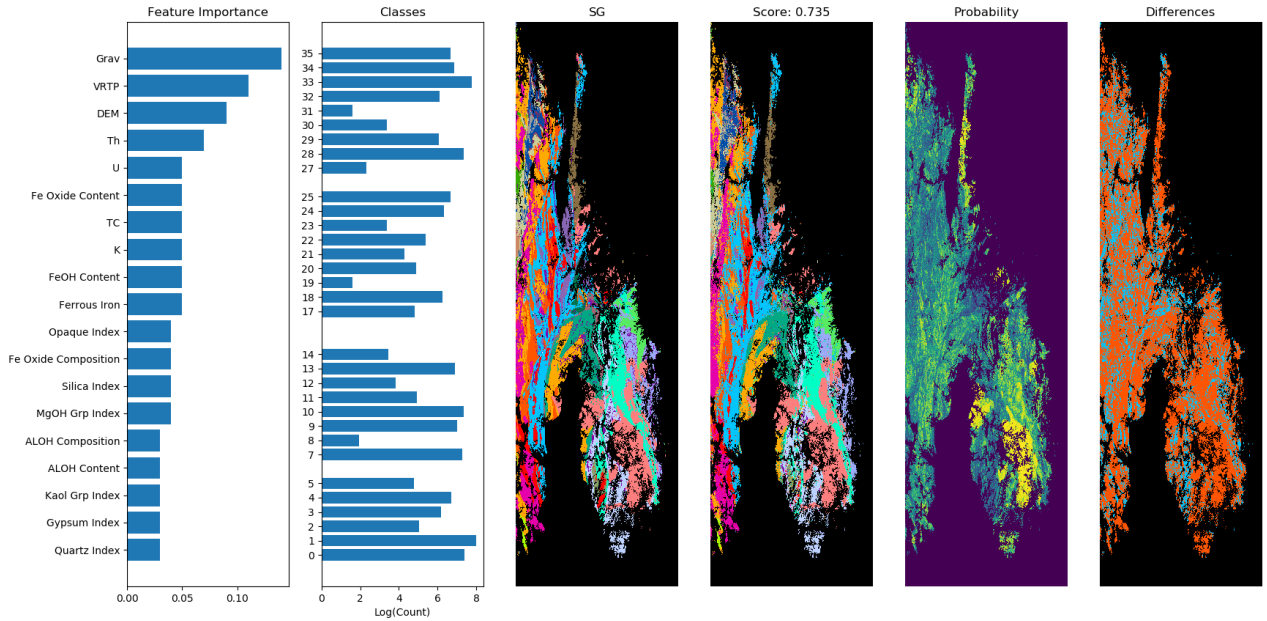



Figure 6.2.: Interpreted geology based prediction (1% samples, all covariates). Column 1 is feature importance, column 2 is the natural log of how many samples where drawn from each class, column 3 is the GSQ supplied simplified surface geology map, column 4 is the RFC estimated surface geology (model accuracy score is shown above the plot), column 5 is the RFC estimated probability and column 6 is Difference between column 3 and 4 (Red = No Change)

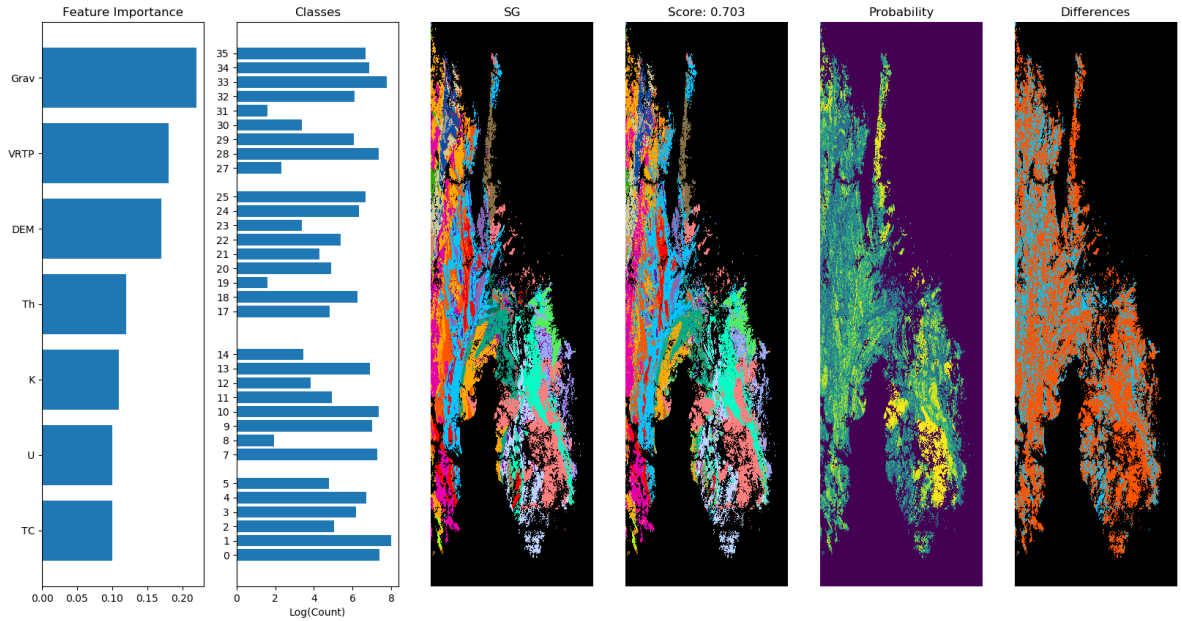



Figure 6.3.: Interpreted geology based prediction (1% samples, no ASTER). Column 1 is feature importance, column 2 is the natural log of how many samples where drawn from each class, column 3 is the GSQ supplied simplified surface geology map, column 4 is the RFC estimated surface geology (the model accuracy score is shown above the plot), column 5 is the RFC estimated probability and column 6 is the difference between column 3 and 4 (Red = No Change)

## Uniform Sampling Including Covariate Range Extents

In most cases, it could be assumed that the number of potential classes in a particular area is not known, while the range of the input covariate data is known. This premise is an important consideration when modelling with the RFC, since the RFC model is not capable of extrapolating outside of the values provided by the training covariates. It is therefore prudent to include locations that correspond to the maximum and minimum of a given covariate so that any value of the covariate thereafter will fall between these two values. The results here were calculated without the use of ASTER products as input covariates.

Figure 6.4 demonstrates that if the maxima and minima of the seven covariates are used along with additional 400 uniformly sampled points from the dataset an accuracy of approximately 50% is achieved and 26 classes are identified showing good correspondence between the modelled and published geology. This result demonstrates that the number of sampling locations used as a basis for training can be quite limited and still yield good results.

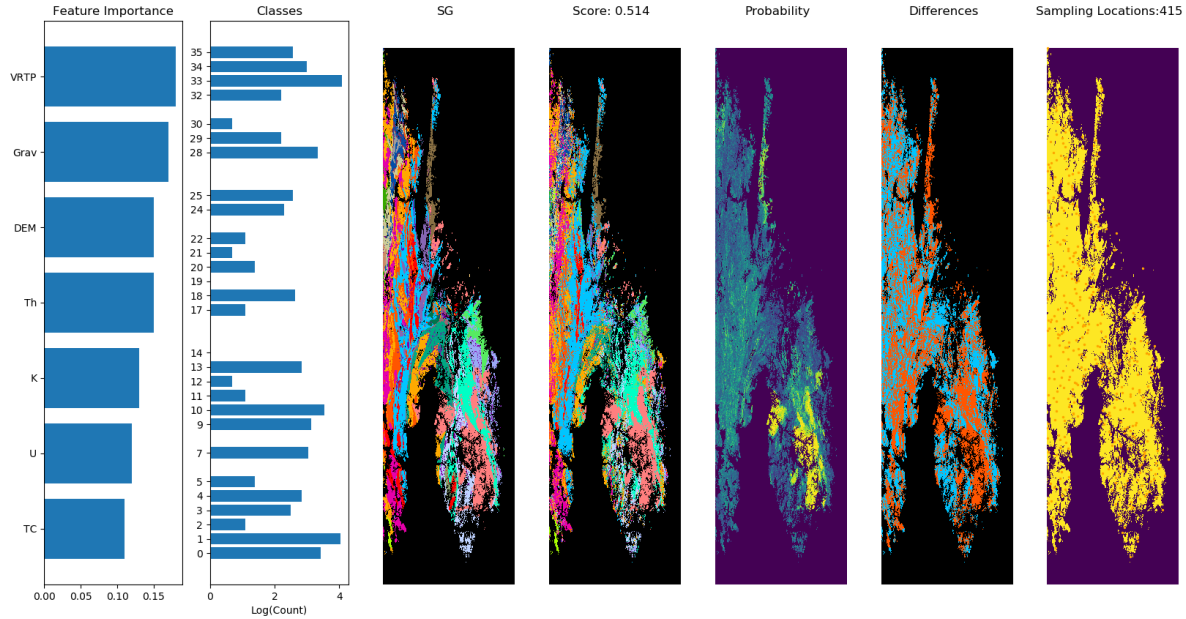



Figure 6.4.: The modelled geology when the training data is restricted to the locations corresponding to the maxima and minima of each of the seven covariates and 400 uniformly selected points. Column 1 is feature importance, column 2 is the natural log of how many samples where drawn from each class, column 3 is the GSQ supplied simplified surface geology, column 4 is the RFC estimated surface geology (Model score is shown above the plot), column 5 is the RFC estimated probability and column 6 is the difference between column 3 and 4 (Red = No Change) and column 7 shows the locations of the samples.

### Conclusions

The scenarios presented in this section successfully demonstrate that if a field operator is able to identify the classes present within a given spatial region, and if ancillary/co-variate data exists at the same locations, it is possible to produce a digital-geology map over the spatial extent under examination with reasonable accuracy using a random forest classifier.

The benefit of the approach is that the feature data, which is assumed to cover the region, is able to be utilised in combination with the point location identifications of the classes such that the actual class boundaries are well resolved. This can be desirable when the actual boundaries of each class are not known. Secondly, and in the absence of total class coverage, it was demonstrated that reasonable geology maps can still be reproduced. The maps produced under such circumstances can be thought of as base level maps whose complexity and accuracy grows as more samples are collected and more classes defined.

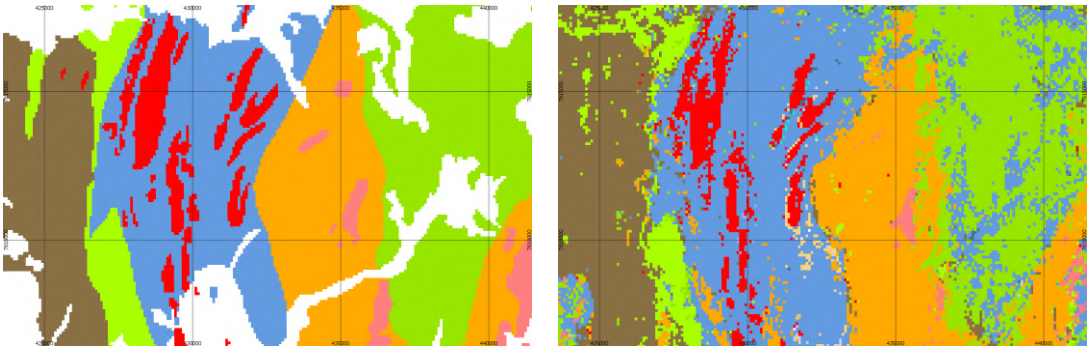
Although not performed in this case, it is noted that a smart sampling method could be utilised to minimise the uncertainty of any class maps produced in this manner allowing the geology model described by the RFC to be updated such that the highest accuracy is achieved.

## 6.1.3. Geological assessment

A geological map is a product of field observations, interpretation of geophysical and remote sensing datasets, as well as expert background knowledge about the geological history of the map area. Traditionally, this process is conducted iteratively, through an integrated visual assessment of various datasets, constrained by field observations, collected across multiple field seasons. We assessed the suitability of a random forest classifier (RFC) in order to provide a repeatable, faster method for interpreting field data and geophysical datasets.

The first described test case samples 10% of simulated field observation in well-exposed terrain (Figure 6.5). The RFC is able to correctly reproduce the geological picture at a similar detail as the GSQ geology map.

The second test case samples 10% simulated field observation in an area with about 50% cover west of the Snake Creek anticline (Figure 6.6). The RFC is able to correctly reproduce the geological picture at a similar detail as the GSQ geology map.

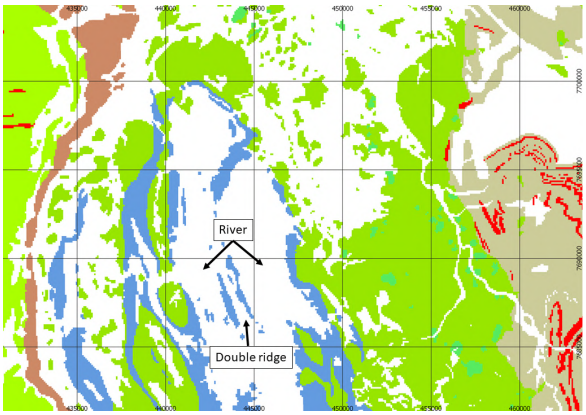


(a) Interpreted geology. (b) Random forest output.

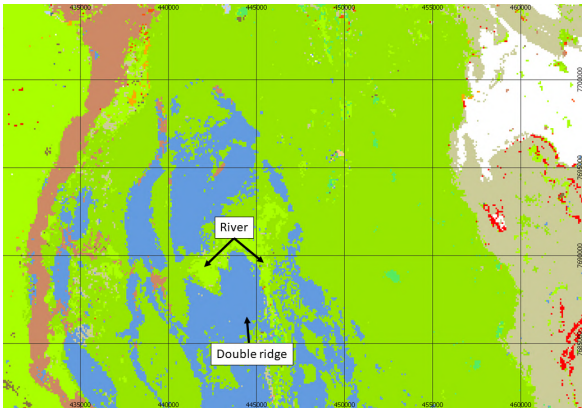
Figure 6.5.: Random forest vs. interpreted geology (example 1).

The third test case simulated a scenario where field observations were gathered from a specific area, and the geology was interpreted outside of that area. In Figure 6.7, the Snake Creek anticline south of Cloncurry has not been sampled. Instead, an area to the north has been sampled with 30 samples from each encountered geological unit. The RFC image shows a geological interpretation that resembles the general distribution of units although cannot provide the detail nor the continuity of those units. It does, however, provide a first pass map that can be utilised for follow-up fieldwork.

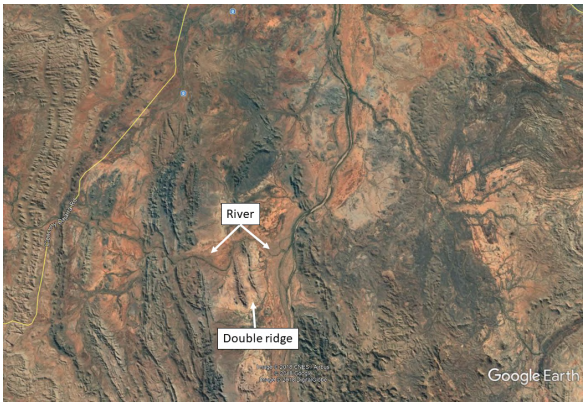
All three examples demonstrate the ability of a RFC algorithm to provide a meaningful automated classification of geological units which could assist in geological mapping and help to guide field work while minimising the time needed to produce a geological map. The described cases also show that in its raw form the automated classifications cannot replace interpreted geological maps. The RFC images display a distinct lack of spatial continuity and a noisy appearance caused by the pixel/point based classification algorithm, as well as misclassifications of geological units probably caused by similar responses of lithologies in the geophysical and remote sensing data. The most significant advantages of the presented approach are: (1) a relatively easy provision for an integrated semi-automated analysis of multiple datasets; (2) the ability to identify potential subtle differences within units; and (3) if used interactively during the mapping process, possibility to increase valuable information gain during fieldwork through guidance to areas of high variability and uncertainty.




(a) Interpreted geology.

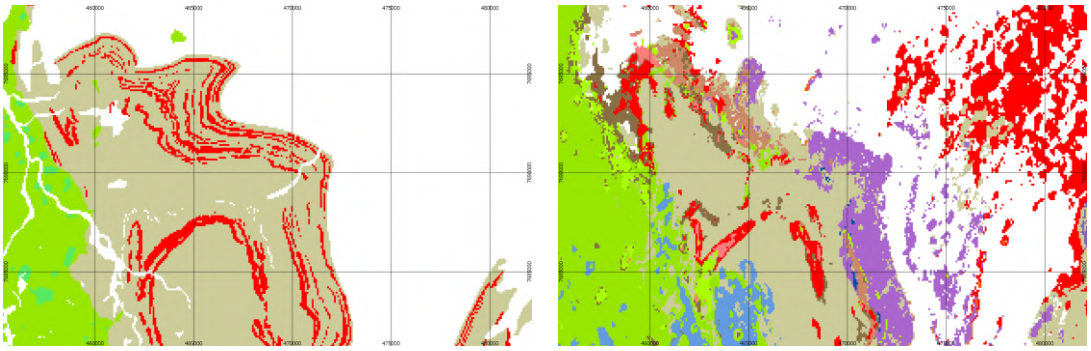



(b) Random forest output.




(c) Satellite imagery.

Figure 6.6.: Random forest vs. interpreted geology (example 2).



(a) Interpreted geology. (b) Random forest output.

Figure 6.7.: Random forest vs. interpreted geology (example 3).

## 6.2. Sparse classification from surface observations

Constructing solid geology maps from surface observations is a time-consuming manual process. Given a relatively sparse set of such observations, it may be possible to utilise a classification algorithm to assist the creation of a solid geology map using available data with continuous regional coverage, such as geophysics and remote sensing. Visual assessment and subjective interpretation of such continuous datasets are routinely used by geologists in the preparation of geological maps. This experiment is concerned with the case where an area has very few surface observations, perhaps only a handful for each rock type under consideration. Semi-supervised learning may improve performance with this few training points, and we will investigate its performance compared to standard supervised classifiers. Semi-supervised classification considers both the target labels but also similarity within the covariate queries. This approach allows us to obtain a classification with only a handful or even a single labelled example of some classes.

### 6.2.1. Problem Formulation

In this experiment, we aim to label every pixel in a multi-band image, given a set of sparse labels for some pixels in that image corresponding to original primary geological

observations. The bands are given by geophysical and remote-sensing covariates, such as gravity, magnetics, and satellite imagery. The targets are categories of 32 different "rock types" defined on the simplified 1:500,000 scale solid geological map produced by GSQ. The assumption in the problem is that there is a relationship between the covariates and these targets. It is this relationship we aim to generalise in order to label all pixels in the image. To simulate circumstances in which very few surface observation exists, we will sub-sample the available surface target data.

### 6.2.2. Target Observations

The target data classified the outcropping lithostratigraphic units into one of 32 classes. The number of available field observations of each class varied substantially, from two sampling points in the Kurbayia Metamorphic Complex to 1515 observations of the Ballara Quartzite/Corella Formation. The full class counts are given in Table 6.1, though for this investigation they were heavily sub-sampled (randomly) in the range from 2 to 200 points per class. In the case where an experiment required more points for a class than were provided, we supplied all available points but used balanced class weights in the classifier.

### 6.2.3. Covariates

We divided the covariates into two categories: geophysical measures of magnetics and gravity, and "surface" remote sensing including ASTER, digital elevation model and gamma-ray spectrometry. This separation allows us to analyse the effects of features sensitive to the surface vs. the subsurface information, and the results when these are combined. The detailed list of covariates is given below.

#### Geophysical

- Gravity
- Magnetics
- Magnetics textures (GLCM)
- Magnetics extracted lines

#### Surface

- ASTER
- DEM
- DEM derived features (TRI, TPI)
- MrVBF
- Radiometrics (gamma-ray spectrometry)

The covariate data were pre-processed for ingestion into the model. Each band was standardised (i.e. offset and scaled) to have zero mean and unit standard deviation. This normalisation tends to assist machine learning algorithms by removing irrelevant differences in scale between the covariates. It is especially important for models that have multiplicative weights which are usually regularised to a single scale over all covariate dimensions.

### 6.2.4. Models

The experiment examined the performance of three models: a semi-supervised approach called label-spreading, and two standard fully-supervised algorithms: random forest and logistic regression. A brief description of each model follows below.

#### Supervised

Supervised classification models use target-covariate pairs to train a model that tries to predict an unknown target from a given covariate. They are supervised in the sense that they are being provided with examples of the relationship they are attempting to learn. For classification problems, the target is a categorical variable with two or more possible values.

#### Random Forest

Similarly to the previous experiment (Section 6.1.1), a random forest classifier (RFC) was again used here. As before, the RFC implementation used in this experiment is from Python's `scikit-learn` package [30].

## Logistic Regression

Logistic Regression [18, Chapter 4.3][29, Chapter 3.2] is a simple linear model for classification, in which the linear latent function is passed through a *softmax*[2] function such that the output is a calibrated probability distribution over the possible target categories. This is a robust but simple algorithm that serves as a good baseline against which to compare other models. In the multi-class setting, it is sometimes referred to as *multinomial logistic regression*. Python's `scikit-learn` package's [30] implementation of a Logistic Regression model was used here.

## Semi-Supervised

Semi-supervised classification is a method for predicting categorical targets from a combination of labelled and unlabelled examples. Algorithms typically try to build decision functions that respect the assignments given in the labelled data but also are "smooth" with respect to the unlabelled data. That is, they exploit the structure of the unlabelled data under the assumption that points nearby in the unlabelled data, or points on the same low-dimensional manifold (e.g., on a line, in a blob or ribbon) should have similar outcomes.

When the query set is known ahead of time, semi-supervised classification can provide additional accuracy compared to a purely supervised approach, especially when the number of labelled examples is small. As such it may be useful for geological classification problems utilising existing geophysical and remote sensing data, where there are too few labelled pixels (observations) to train a traditional classifier.

## Label Spreading

In this experiment, our semi-supervised method was the "Label Spreading" algorithm in the Python `scikit-learn` package [30]. Label Spreading [31] is a variation on the Label Propagation algorithm which propagates labels from *labelled* samples to *unlabelled* samples which are near-by in the feature space.

Label spreading envisions the pixels in the image as being points in a high-dimensional space and constructs a graph by connecting nearby points. The distance metric used is

[2] The *softmax*, or *normalised exponential*, function is a multi-dimensional generalisation of the sigmoid function, used here to map an $D$-dimensional continuous variable to a multinomial probability distribution over $D$ classes.

configurable, but in this case, we used a $k$-nearest-neighbour algorithm and an L2 (i.e. Euclidean) distance measure over the covariates.

Once the algorithm has constructed this graph, it initialises nodes with known labels. Then, the algorithm iteratively propagates information about labels to nearby nodes. Ultimately the label of each pixel is determined by what information it has received about nearby labels, as well as its original label. This algorithm therefore includes the possibility of mislabelled (i.e. noisy) pixels. Label spreading also includes a "regularisation" term, which tries to provide a "smooth" labelling that does not have too much local variation.

## 6.2.5. Results

### Cross-Validation Performance

We compared the performance of the three models as a function of the number of examples for each of the 32 classes. We compared performance using $k$-fold cross-validation. The number of folds was 5 for all runs except those with fewer than 5 points, in which case a leave-one-out cross-validation was used. The results are provided in Figure 6.8. The accuracy measure used is the average accuracy over all the classes (balanced). Note that in a 32-class problem, "random" accuracy is not 0.5 but 1/32 (i.e. 0.03). Therefore it is significant that even with 2 points, all models make predictions significantly better than random.

The label-spreading algorithm under-performed compared to the others, even with only 2 training points and when given additional 50k unlabelled points. The random forest and linear models perform similarly when the number of training points is low, but the additional expressiveness of the random forest allows it to increase performance as the number of training points increases. The failure of the semi-supervised method is probably due to two factors: the high dimensionality of the covariate space, and the lack of a low-dimensional manifold in the covariate space (i.e. there is no simple underlying structure in the dataset which can be exploited to infer the classes of unlabelled data). As such the unlabelled pixels did not contribute (and appeared to mostly hinder) the method.

Based on the cross-validation accuracy, the surface features appear to have more relevance for the lithostratigraphic classes than the geophysical features. This would be expected, given both are measurements of surface properties. However, the combination

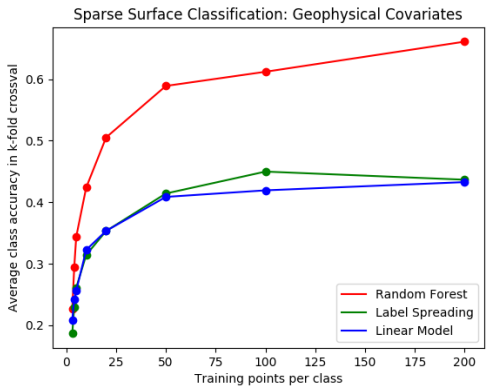
of both surface and geophysical features yielded the best results, indicating these two sets of covariates are encoding different information about the targets.

## Geophysical and surface features

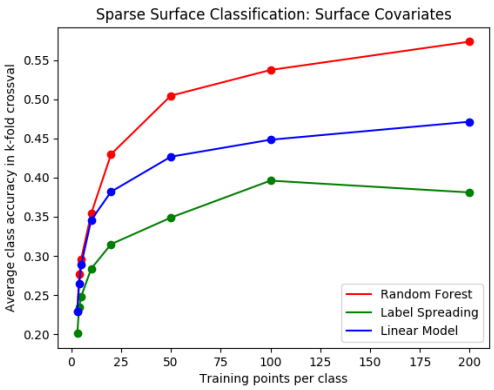
The results of the combined geophysics and surface remote sensing information showed the best results. Geologically, these combined outcomes present the most useful information. Nevertheless, the scarcity of locations of primary geological observations and their uneven spatial distribution probably causes the classifier to perform with lower accuracy than in the case of the more evenly spread sampling points randomly selected from the interpreted geology. In the north, we can see that Ballara Quartzite and Corella Formation are spatially underestimated and confused with Boomarra Metamorphics, Milo Beds, Leichhardt Volcanics and Argylla Formation/Magna Lynn Metabasalt. This can be caused by the fact that the Boomarra Metamorphics have a similar estimated magnetic susceptibility ~30-40 $\times 10^{-3}$ SI to the Ballara Quartzite/Corella Fm unit and also by the relatively lower thickness of the Corella formation [15]. Interesting is also the complete confusion of the Mount Albert Group with the Toole Creek Volcanics and Milo Beds. All of these units show low magnetic susceptibility ~0-2 $\times 10^{-3}$ SI [15], with the Milo Beds at the higher end of the range. The Mount Albert Group has two magnetically distinct domains. In the northern part, the unit displays no texture while the southern end near the outcrop of the Milo beds shows moderate textural response and overall higher susceptibility. This bimodal character could be one of the reasons for the misclassification, further enhanced by the fact that the field observations, used for training, were made mostly over the more magnetic part of the unit.

Table 6.1.: Sparse geological classification class observation counts.

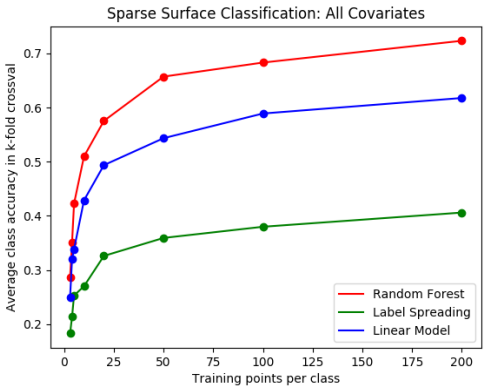
| Class | Symbol | Name | Measurement sample count (total) |
|---|---|---|---|
| 0 | PLea | Argylla Formation, Magna Lynn Metabasalt | 873 |
| 1 | PLk | Ballara Quartzite, Corella Formation | 1515 |
| 2 | PLf | Bigie Formation, Fiery Creek Volcanics | 23 |
| 3 | PLbb | Boomarra Metamorphics | 252 |
| 4 | PLeb | Bulonga Volcanics | 282 |
| 5 | PLdc | Double Crossing Metamorphics | 275 |
| 6 | PLgk | Kalkadoon Supersuite | 207 |
| 7 | PLlk | Kurbayia Metamorphic Complex | 2 |
| 8 | PLkr | Kuridala Group | 720 |
| 9 | PLel | Leichhardt Volcanics | 372 |
| 10 | PLh | Lower Haslingden Group | 57 |
| 11 | PLn | Malbon Group | 337 |
| 12 | PLgm | Maramungee Suite | 8 |
| 13 | PLm | Milo beds | 211 |
| 14 | PLp | Mount Albert Group | 534 |
| 15 | PLmfc | Mount Fort Constantine Volcanics | 7 |
| 16 | PLi | Mount Isa Group | 21 |
| 17 | PLhm | Myally Subgroup | 16 |
| 18 | PLkj | Overhang Jaspilite | 278 |
| 19 | PLuq | Quamby Conglomerate | 12 |
| 20 | PLq | Quilalar Formation | 149 |
| 21 | PLo | Soldiers Cap Group | 552 |
| 22 | PLo? | Soldiers Cap Group? | 1 |
| 23 | PLxs | Stanbroke Sandstone | 6 |
| 24 | PLks | Staveley Formation, Roxmere Quartzite | 1114 |
| 25 | PLr | Surprise Creek Formation | 85 |
| 26 | PLgt | Tommy Creek Microgranite | 41 |
| 27 | | Too small | 7 |
| 28 | PLot | Toole Creek Volcanics | 195 |
| 29 | PLgi | Williams Supersuite | 390 |
| 30 | PLgw | Wonga Suite, Burstall Suite | 227 |
| 31 | Plkr | do-MI | 201 |




(a) Geophysical covariates.

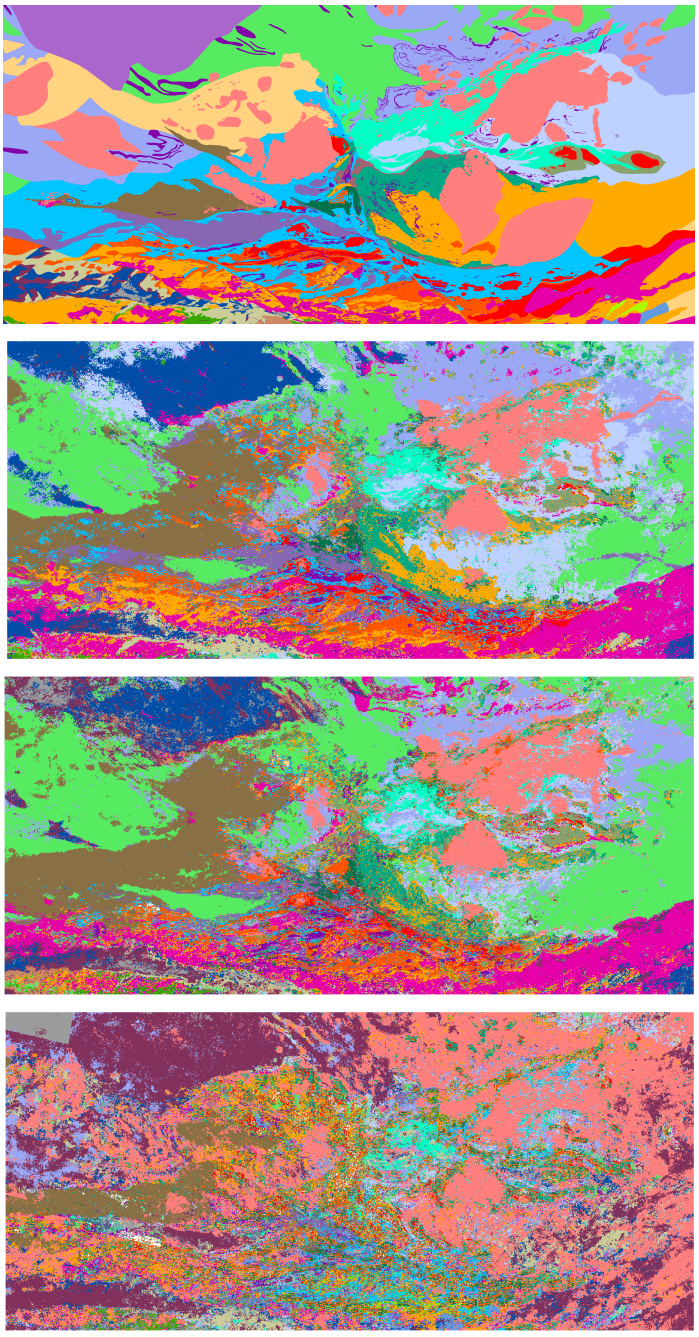



(b) Surface covariates.




(c) All covariates.

Figure 6.8.: Sparse geological classification cross-validation accuracy. Each graph shows how the model accuracy changed as a function of the number of labelled training points per class. The analysis was run for three different sets of input (covariate) data.



(a) Model prediction – 2 points per class.

(b) Model prediction – 20 points per class.

(c) Model prediction – 200 points per class.

(d) Solid geology interpretation.

Figure 6.9.: Sparse geology classification – *Random forest* with *all* covariates. The model prediction is shown for varying number of training points (a maximum of 2, 20, and 200) per class. Figure 6.9d shows the existing solid geology interpretation for comparison.

## Geophysical features only

The classifications based on the magnetic and gravity data only (Figure 6.10) are not too dissimilar from the combined classification results, indicating that the information depicting the below-cover geological features plays a significant role in the classification. This has been previously confirmed by the classification based on training data acquired from the interpreted geology. One of the main differences is the breakup of the fairly coherent Williams and Maramungee Suite granites and higher confusion with the surrounding lithologies. The less accurate mapping of the granitic bodies might be the effect of the missing gamma-ray spectrometry data, which define the granites more clearly, at least in the outcropping parts of the region.

## Surface features only

As in the previous examples, one can see that more than 20 samples per class are needed to obtain a reasonable and more consistent regional prediction (Figure 6.11). In this classification result, we can clearly observe the effect of the surface data on the final prediction with the appearance of drainage patterns, possibly indicating the source of the sediment material. The classification does a relatively good job over the outcropping regions recognising even smaller granitic bodies but completely falls apart in the covered regions. We can also observe a higher fragmentation of the classes probably due to the inherently noisier nature of the gamma-ray data and the overall higher spatial resolution of the input variables.

## Semi-supervised predictions

The semi-supervised predictions are constructed differently from the fully supervised ones, and so may have a different geological interpretation (especially in regions away from the training points). The label spreading classifications produce the most coherent results, showing good correspondence to the map, in areas that are relatively close to the site observations. With increasing distance, the accuracy seems to drop. As in previous examples, some units are completely misclassified, Mount Albert Group for example, while some like the Williams and Maramungee Suite granites can be recognised relatively well at least in the southern part of the study area. Very well noticeable is an artefact in the north-eastern corner derived from the ASTER data probably caused by issues with the mosaicking process.

(a) Model prediction – 2 points per class.

(b) Model prediction – 20 points per class.

(c) Model prediction – 200 points per class.

(d) Solid geology with RTP shading.

Figure 6.10.: Sparse geology classification – *Random forest* using only *gravity and magnetics* covariates. The model prediction is shown for varying number of training points (a maximum of 2, 20, and 200) per class. Figure 6.10d shows the existing solid geology interpretation with magnetic RTP shading for comparison.

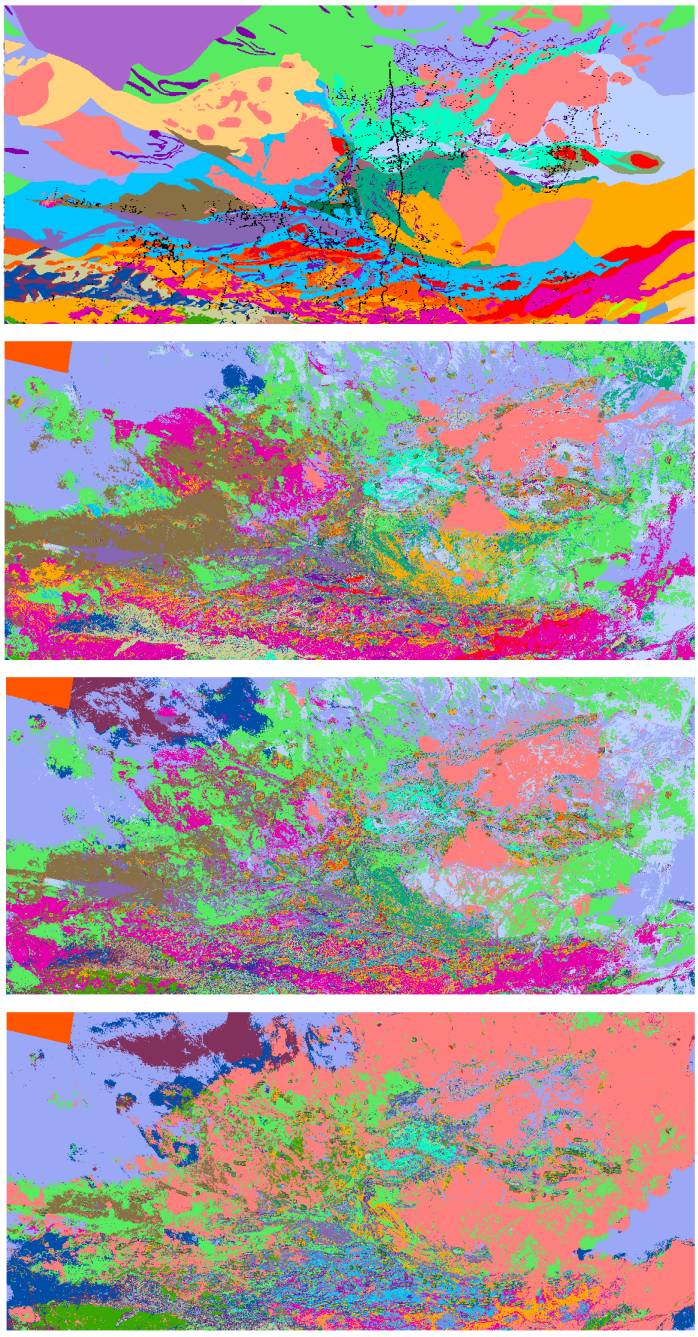
(a) Model prediction – 2 points per class.

(b) Model prediction – 20 points per class.

(c) Model prediction – 200 points per class.

(d) Gamma-ray spectrometry draped over 1VD TMI.

Figure 6.11.: Sparse geology classification – *Random forest* using only *remote sensing* covariates. The model prediction is shown for varying number of training points (a maximum of 2, 20, and 200) per class. Figure 6.11d shows the gamma-ray spectrometry draped over 1VD TMI for comparison.



(a) Model prediction – 2 points per class.

(b) Model prediction – 20 points per class.

(c) Model prediction – 200 points per class.

(d) Solid geology interpretation with training points.

Figure 6.12.: Sparse geology classification – *Label spreading* with *all* covariates. The model prediction is shown for varying number of training points (a maximum of 2, 20, and 200) per class. Figure 6.12d shows the existing solid geology interpretation with training points overlaid for comparison.

### 6.2.6. Geological assessment

The sparse observation problem corresponds to the classical situation of traditional geological mapping, where limited surface observations can only be made to produce an interpreted geological map. In this scenario, the placement of the training samples is crucial, and we can see a detrimental effect of the uneven limited number of observations on all of the classification attempts. A relatively high number of samples does not automatically guarantee a good classification result. We can probably predict that well-placed training data/observations could produce a good starting point for a more accurate interpretation. Some of the mismatches that were identified indeed deserve a more thorough assessment and re-evaluation of the mapped geology. The results also show that a combination of both the deeper imaging geophysical data and the surface mapping geophysical and remote sensing data produces the best results. Integration of all data only visually tends to be a difficult task, and the illustrated method might significantly aid a mapping geologist as a first pass prediction of the distribution of lithological units.

# 7. Classification of outcropping granites

This section describes a model for classifying outcropping granites based on geospatial data. A random forest classifier was trained on labelled data consisting of the 5 granite classes present (PLgk, PLgm, PLgt, PLgi, PLgw) with all other rock types labelled as "None". The trained model was then used to predict granite classes across a large section of the eastern area of the NWMP, and the output compared to the interpreted geology map. The results show general agreement with the interpreted geology with higher accuracy for the more spatially extensive granite types, while the less frequently occurring granite types were difficult to predict. The model has potential to be used as a starting point for complete interpreted granite map, and disparities between the model and the existing interpreted geology may indicate areas which require further investigation.

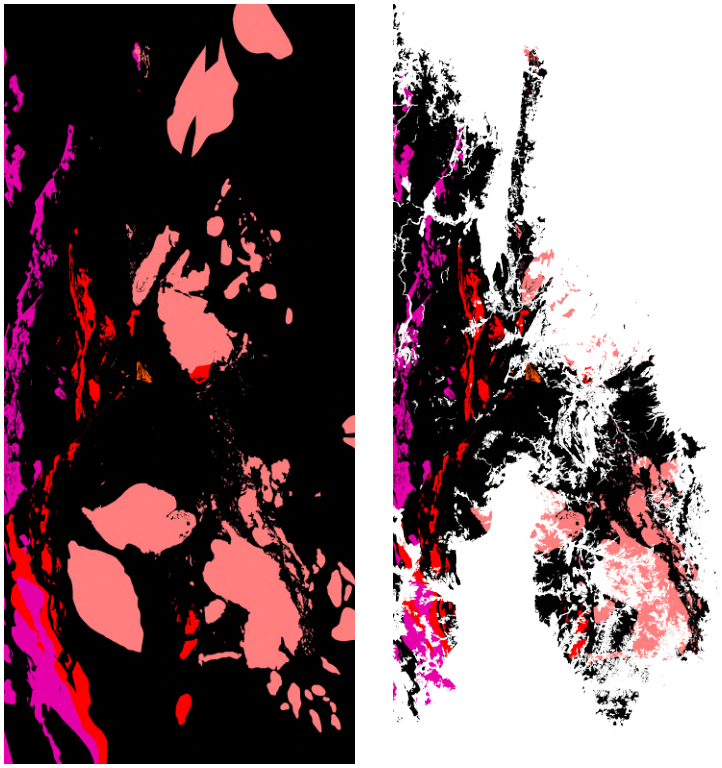
## 7.1. Model inputs

Models were trained separately with target training data from two different sources:

1. ~6,000 Surface observations
2. 50,000 random points from the outcropping area with classes chosen from the interpreted geology

The above training sample sets cover 0.3% and 2.4% respectively of the rasterised interpreted geology map used to validate the model.

The various covariates used are shown in Table 7.1, grouped in order to reference the combinations they were used in for each model. The combination of covariates considered for the randomly selected training data are grouped as follows: surface data (A); surface data and geophysics (A + B); and surface data, geophysics, and mineral maps (A + B + C). The mineral map data (C) was considered separately in order to



(a) All granites. (b) Outcropping granites.

Figure 7.1.: Granite classes identified in the solid geology map.

quantify how much additional information it contains relative to this prediction task.

Table 7.1.: Covariate datasets for granite classification.

| A | B | C |
|---|---|---|
| ASTER | Gravity | Mineral maps |
| DEM + TRI/TPI | Magnetics 1VD + GLCM textures | |
| MrVBF | Magnetics VRTP + GLCM textures | |
| Radiometrics (Gamma-ray spectrometry) | + distance to lines | |

## 7.2. Model results

The resulting classification performance (precision, recall, and $F_1$ score) [29, Chapter 5.6] of the predicted granite map (minus any training point locations) vs. the interpreted granite map are shown for each granite class in Tables 7.2-7.5 below. The prediction maps and uncertainties for the surface observation model (model 1) and the random points (covariates A + B) model (model 3) are shown in Figures 7.2 and 7.3.

The classifiers are reasonably accurate when identifying all granite types (high precision), yet for some of the smaller granite intrusions the classifiers are not detecting a large proportion of the bodies (low recall). The overall performance (captured by the $F_1$ score) for the major granite (PLgi) and no granite areas (None) is very good, and for the next two largest granite types (PLgk, PLgw) it is reasonable for the random training points and A + B covariates.

The classifiers using training points selected randomly from the interpreted geology perform much better than those with the surface observation training set. This is mostly due to the increased number of training points.

By considering the results shown in Tables 7.4 and 7.5 below, it can be seen that any improvement in performance with the addition of the mineral maps as a covariate is marginal. The addition of the magnetics and gravity data does, however, improve the performance significantly over the surface data only model (Table 7.3).

As the complete interpreted geology map is being used to measure the performance of the models, the errors indicated in the cross-validation tables and the figures below

Table 7.2.: Model 1 performance: Surface observations / Covariates A + B

| | **precision** | **recall** | $F_1$ **score** | **support** |
|---|---|---|---|---|
| None | 0.86 | 0.98 | 0.92 | 1568834 |
| PLgk | 0.69 | 0.21 | 0.32 | 142288 |
| PLgm | 1 | 0 | 0 | 3259 |
| PLgt | 0.91 | 0.08 | 0.14 | 2884 |
| PLgi | 0.86 | 0.74 | 0.8 | 235169 |
| PLgw | 0.84 | 0.1 | 0.18 | 95336 |
| **avg/total** | **0.85** | **0.86** | **0.83** | **2047770** |

Table 7.3.: Model 2 performance: Random points / Covariates A

| | **precision** | **recall** | $F_1$ **score** | **support** |
|---|---|---|---|---|
| None | 0.88 | 0.98 | 0.93 | 1536356 |
| PLgk | 0.77 | 0.27 | 0.4 | 139030 |
| PLgm | 0.93 | 0 | 0.01 | 3189 |
| PLgt | 0.5 | 0 | 0 | 2855 |
| PLgi | 0.86 | 0.82 | 0.84 | 229928 |
| PLgw | 0.78 | 0.18 | 0.29 | 93227 |
| **avg/total** | **0.86** | **0.87** | **0.85** | **2004585** |

can be attributed to either misclassifications in the model, or misclassifications in the interpreted geology. The purpose of the model is to highlight areas of disagreement between the two, which may require further attention to the data or a visit to the site, to determine the correct classification. Mismatched regions are highlighted in Figure 7.2. They confirm the precision/recall results showing that the major granite type (PLgi) is most accurately classified and more mismatches appear along the western side among the PLgk and PLgw granite bodies.

A particular area of detail where a mismatch occurs is shown in Figure 7.4. The section in the centre of the image is classified as granite in the interpreted geology but

Table 7.4.: Model 3 performance: Random points / Covariates A + B

| | **precision** | **recall** | $F_1$ **score** | **support** |
|---|---|---|---|---|
| None | 0.91 | 0.98 | 0.94 | 1536356 |
| PLgk | 0.82 | 0.58 | 0.68 | 139030 |
| PLgm | 1 | 0 | 0.01 | 3189 |
| PLgt | 1 | 0.01 | 0.01 | 2855 |
| PLgi | 0.91 | 0.84 | 0.87 | 229928 |
| PLgw | 0.9 | 0.42 | 0.57 | 93227 |
| **avg/total** | **0.9** | **0.91** | **0.9** | **2004585** |

Table 7.5.: Model 4 performance: Random points / Covariates A + B + C

| | **precision** | **recall** | $F_1$ **score** | **support** |
|---|---|---|---|---|
| None | 0.91 | 0.98 | 0.94 | 1536356 |
| PLgk | 0.82 | 0.6 | 0.69 | 139030 |
| PLgm | 1 | 0.01 | 0.02 | 3189 |
| PLgt | 1 | 0 | 0.01 | 2855 |
| PLgi | 0.91 | 0.83 | 0.87 | 229928 |
| PLgw | 0.89 | 0.42 | 0.57 | 93227 |
| **avg/total** | **0.9** | **0.91** | **0.9** | **2004585** |

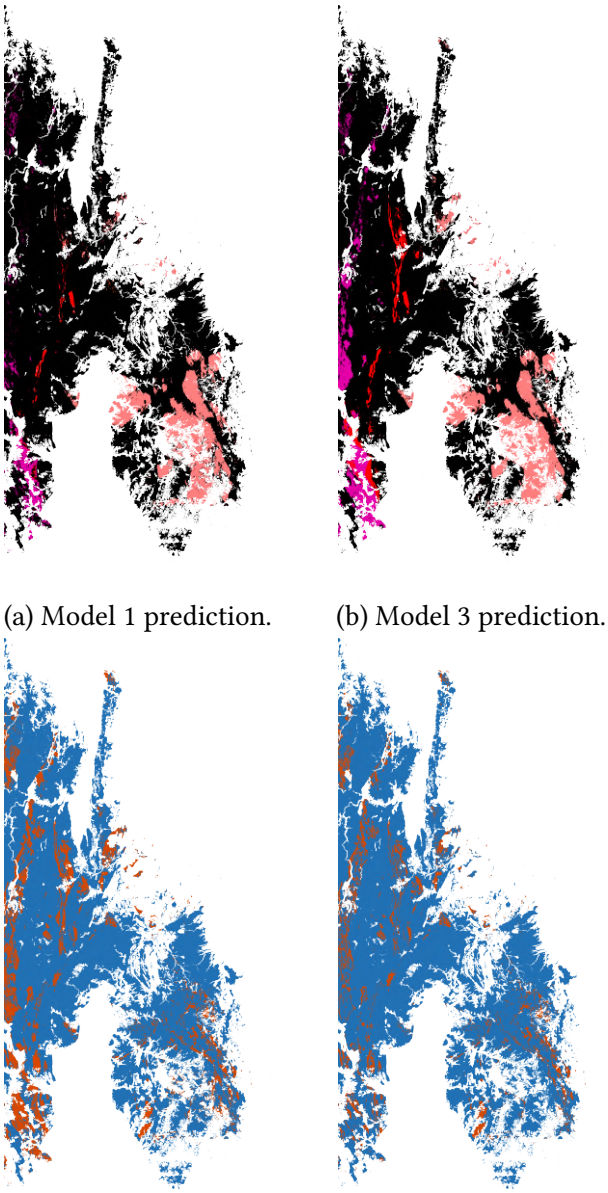
not-granite by the model. The model probability for mismatched areas is shown also, the bright green and yellow areas show much higher model certainty, including the in the central area where the model prediction disagrees with the interpreted geology.

## 7.3. Geological interpretation

The primary purpose of this exercise was to evaluate the possibility of discovering unnoticed granitic rocks that might potentially be associated with mineralised areas. The high accuracy obtained for the outcropping regions shows promising results. The

mismatched area in the central part of Figure 7.4 should be revisited, as neither the geophysical data nor the remote sensing data point to a different signature when compared to the main granitic body. There is a chance that a thin quaternary layer of alluvial sediments might be present and observed in the field or interpreted in airborne photographs. Such a feature would be too subtle to be recognised at the scale of the input data. One of the smaller granitic stocks in the north-central part of the figure also shows a significant mismatch. In this area, a Williams suite granitic pluton was mapped thanks to several surface observations and a rounded shape in the magnetic data, however other surface observations of Toole Creek Volcanics located in the middle of the granitic body and also the gamma-ray spectrometry data show signatures more resembling the Toole Creek Volcanics.

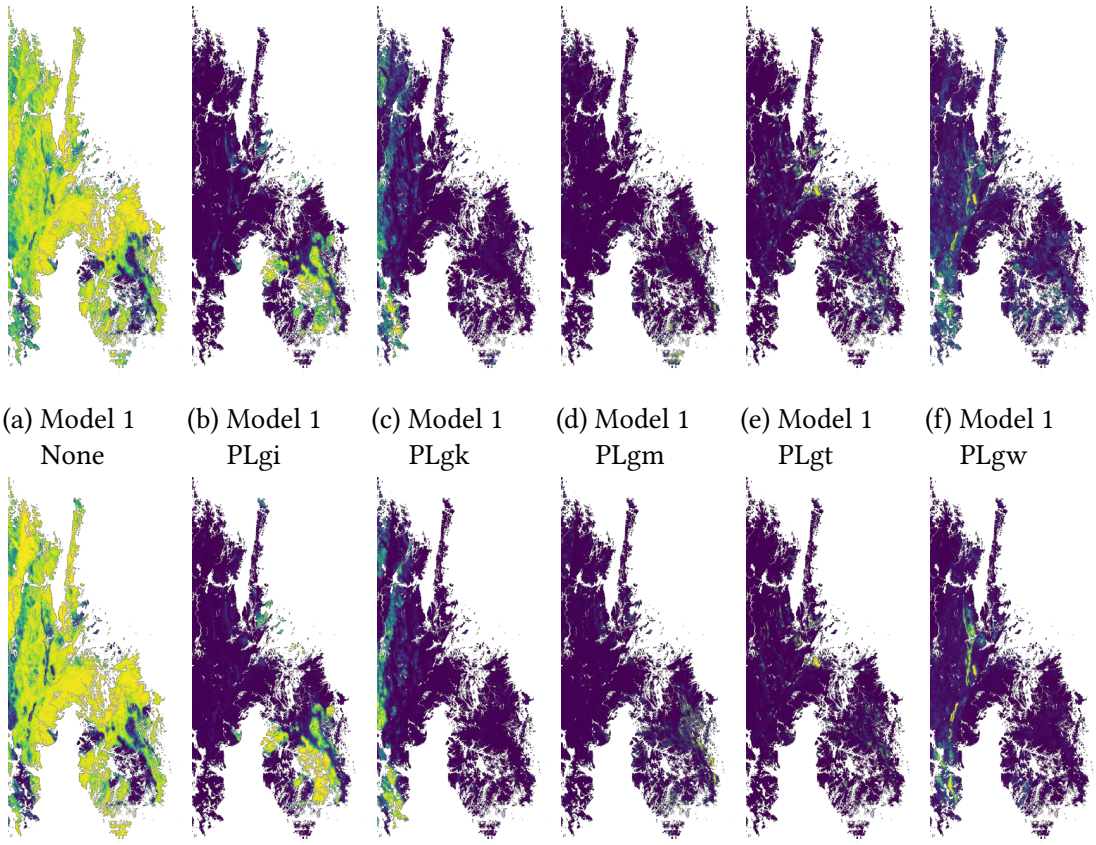
The prediction of granitic bodies in covered areas (Figure 7.5) shows several locations which could be investigated. Apart from some definite drainage-related false positives in the south of the study area, we can observe indications of the mostly covered granitic plutons on the Quamby map sheet in the centre of the map. There are also several small granitic occurrences all along the eastern boundary of the outcropping lithologies which seem to correspond with the interpreted solid geology and point to locations where outcrop could exist, or the intrusive body might be close to the surface.



(a) Model 1 prediction. (b) Model 3 prediction.

(c) Model 1 difference. (d) Model 3 difference.

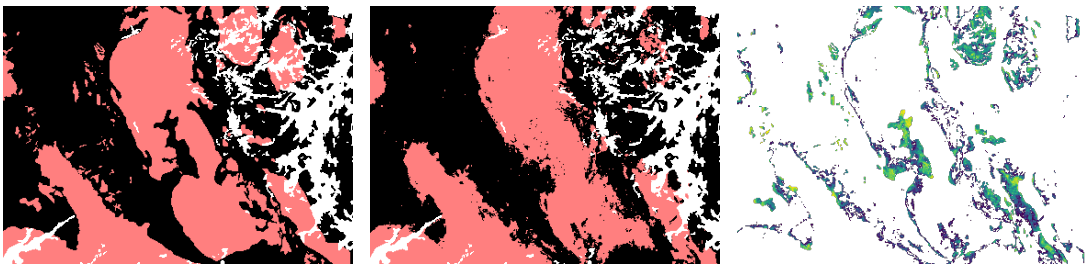
Figure 7.2.: Granite model comparison to interpreted geology. Above: predicted granite distribution. Below: blue areas indicate agreement between the model predicted class and the interpreted class; differences are shown in red.



(a) Model 1 None (b) Model 1 PLgi (c) Model 1 PLgk (d) Model 1 PLgm (e) Model 1 PLgt (f) Model 1 PLgw

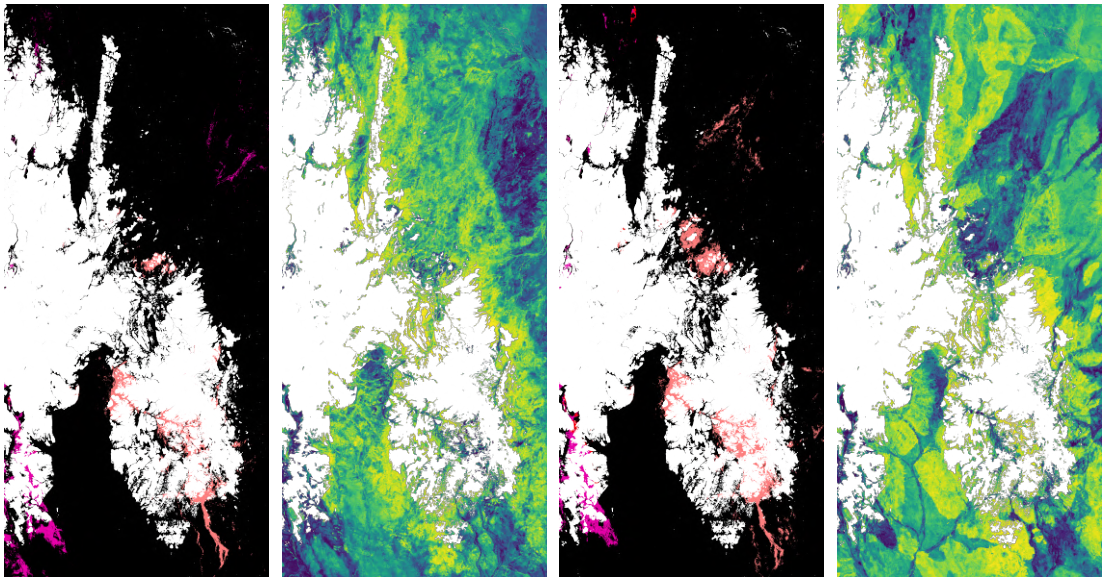
(g) Model 3 None (h) Model 3 PLgi (i) Model 3 PLgk (j) Model 3 PLgm (k) Model 3 PLgt (l) Model 3 PLgw

Figure 7.3.: Individual granite class probabilities for model 1 (Surface observations / Covariates A + B) and model 3 (random points / covariates A + B). Probability estimates range from purple (very low) to bright yellow (high probability). Green areas in between indicate that the model is uncertain in the prediction.



(a) detail interpreted geology. (b) Detail model prediction. (c) Model probability.

Figure 7.4.: Detail of interpreted geology mismatch with model 3 (random points, covariates A + B) prediction. Figure 7.4c shows the models estimated probability for mismatched areas. Bright yellow areas indicate that the model is very confident in the prediction.



(a) Model 1 prediction. (b) Model 1 probability. (c) Model 3 prediction. (d) Model 3 probability.

Figure 7.5.: Granite model predictions in "covered" areas of the interpreted geology. Bright yellow areas within the probability map indicate that the model is very confident in the prediction, based on the training data.

# 8. Interpreted geology anomaly detection

GSQ have developed a simplified interpretation of the solid geology across the Mt Isa east area (Figure 2.1). This interpretation was produced by visual analysis based on all the available geospatial data (e.g., magnetics, gravity, DEM, and ASTER), accounting for surface observations and any exploration drill hole data available. This section describes a machine learning approach to finding areas within the interpreted geology classification which appear to be inconsistent with the data.

## 8.1. Model

A one-class Support Vector Machine (SVM) (see [29, Chapter 7.3] and [30, *OneClassSVM*]) is used to model anomalies within each interpreted lithological class (as defined in the simplified solid geology of the so-called Eastern Succession) according to the magnetics and gravity geophysical datasets.

Of the geospatial data available, only gravity and magnetics are able to provide information below cover. Therefore, in order to produce an output across the entire area (both covered and uncovered), the inputs to the model are limited to these two datasets.

In a visual analysis of the geophysical data, the textural properties are fundamental in identifying geological characteristics of a region and commonalities between regions. To enable the algorithm to take this into account, textural properties (GLCM texture features) of the 1VD magnetics data were derived and used as additional inputs to the model (see Appendix A).

This one-class SVM model identifies the common characteristics of the geophysical data within each rock type by defining a hyperplane — a boundary defined in the high-dimensional feature space — which encloses the bulk of that data. Anomalies are

then identified as points which lie outside the hyperplane, and the distance from it is used as an anomaly measure. That is, the further a point is away from the main group of points of a given class, the more anomalous it is.

The model requires that an estimate of the upper bound on the percentage of anomalous data points be provided. As this value is not known directly, the algorithm was run for multiple values in order to observe the trends in identified anomalies.

A separate one-class SVM model is generated for each of the 36 classes (rock types) present in the solid geology map, and anomalies identified within each class. The *nu* value is an upper bound on the estimated number of points which are anomalous (e.g., $nu = 0.5$ means that up to 50% of the data may be anomalous). The models were run for *nu* values of 0.5, 0.1, 0.05 and 0.01.

### Model inputs:

- Gravity
- Magnetics 1VD
- Magnetics 1VD GLCM textural features
- Magnetics RTP

### Model outputs:

- Binary prediction ("in-class" or "anomalous")
- Distance to hyperplane

## 8.2. Estimated anomalies

The outputs of the models are shown in Figures 8.1-8.3 via three values which aim to highlight the areas which may be inconsistently classified and the geological classes for which there is a higher potential for these inconsistencies:

1. **Anomaly prediction** – Purple indicates the area is anomalous within its class, according to the model. As described above this means the values of the model inputs (geophysical data) for these purple areas differ significantly from the majority of the values for the assigned class (i.e. they lie outside the hyperplane which encloses the majority of the data), therefore they are anomalies with respect to that data.

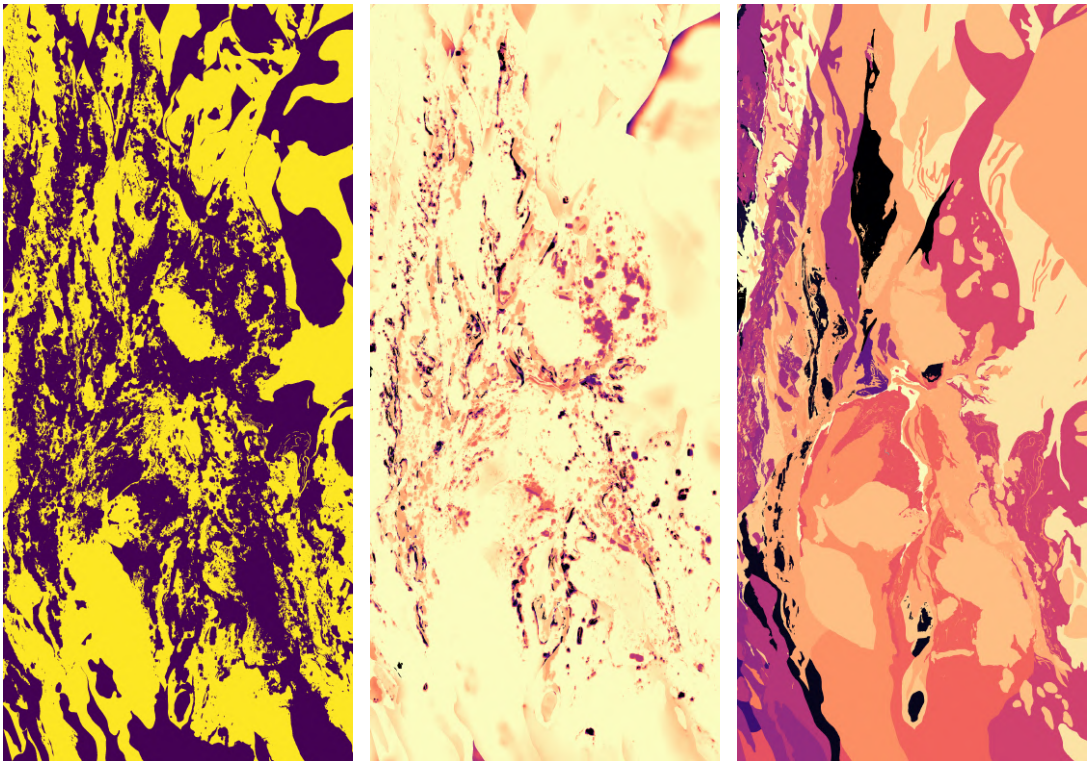
2. **Distance to hyperplane for anomalies only** – A measure of how anomalous a given point is with respect to its class. Light areas are within the class. The anomalous areas are shaded according to how far the point is from the model hyperplane of the class boundary. Darker points are more anomalous.

3. **Mean anomalous distance** – The mean normalised distance for each geology class is presented here to indicate which classes have the most significant anomalies. It is a single value per class, which indicates on average how far away anomalous points are from that class boundary. Darker represents more anomalous.

As the estimated percentage of anomalous points decreases ($nu = 0.5 \rightarrow 0.1 \rightarrow 0.01$) the predicted anomalies tend to focus on the border areas of the classes. This is highlighted by the fact that the geological class with the highest mean normalised anomalous distance is one with a large perimeter to surface area ratio.

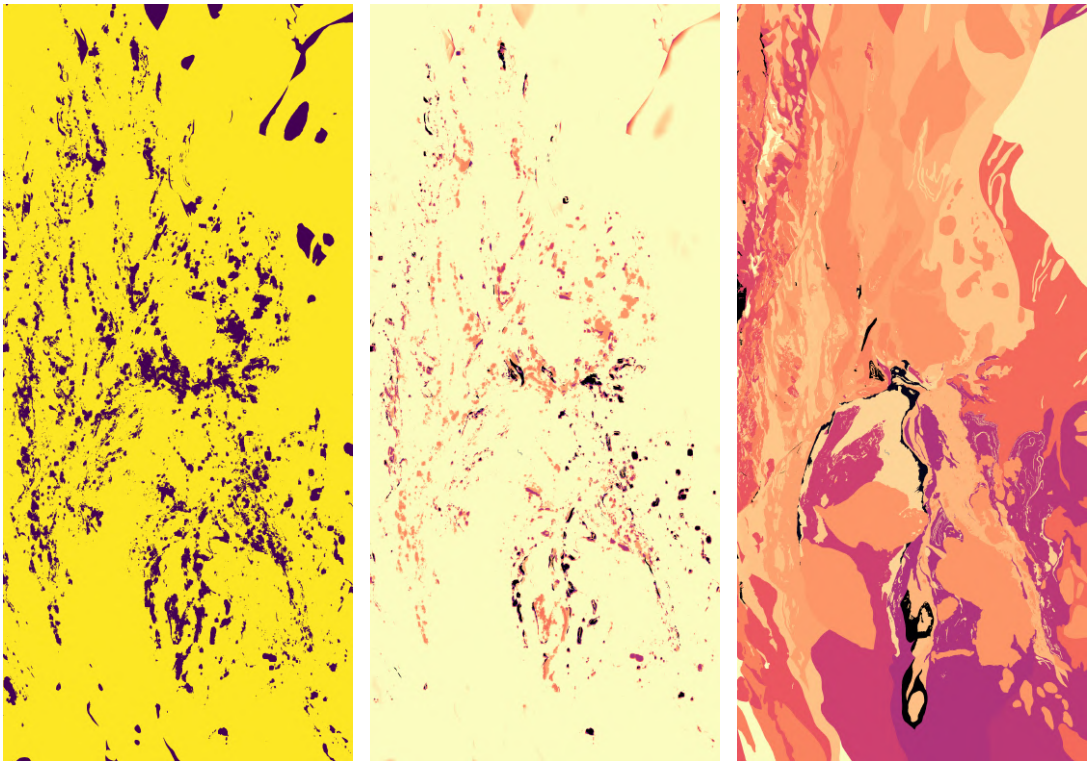
## 8.3. Geological interpretation

Of most use to the geologist will probably be both the results with the most and least estimated amount of anomalous points, as they provide more detail of the variability for the highly anomalous regions and less anomalous regions, respectively. The identification of lithological boundaries as highly anomalous areas is to be expected and only confirms the fuzzy nature of some of these and the uncertainty in their position on the interpreted map. The mean normalised anomalous differences clearly show some of the units that pose problems even in the other classification scenarios, for example, Mount Albert Group, Boomarra Metamorphics, Mount Fort Constantine Volcanics. After identifying the problematic units, one can zoom in on these to identify anomalies that need further assessment. Interesting is the relatively non-anomalous nature of the Williams Suite granitic plutons, even though most of the bodies in the northern part of the study area were misclassified during the other conducted classification experiments. This might be caused by the fact that we pool the result for all the Williams Suite polygons into one average number and favour the not covered areas due to the larger spatial extent, a per-polygon value that would separate the predictions from the covered part of the study region would probably provide a different view.



(a) Predicted anomalies. (b) Anomalous distance. (c) Mean anomalous distance.

Figure 8.1.: Predicted anomalies in the interpreted geological map ($nu = 0.5$). See description in Section 8.2 for details.

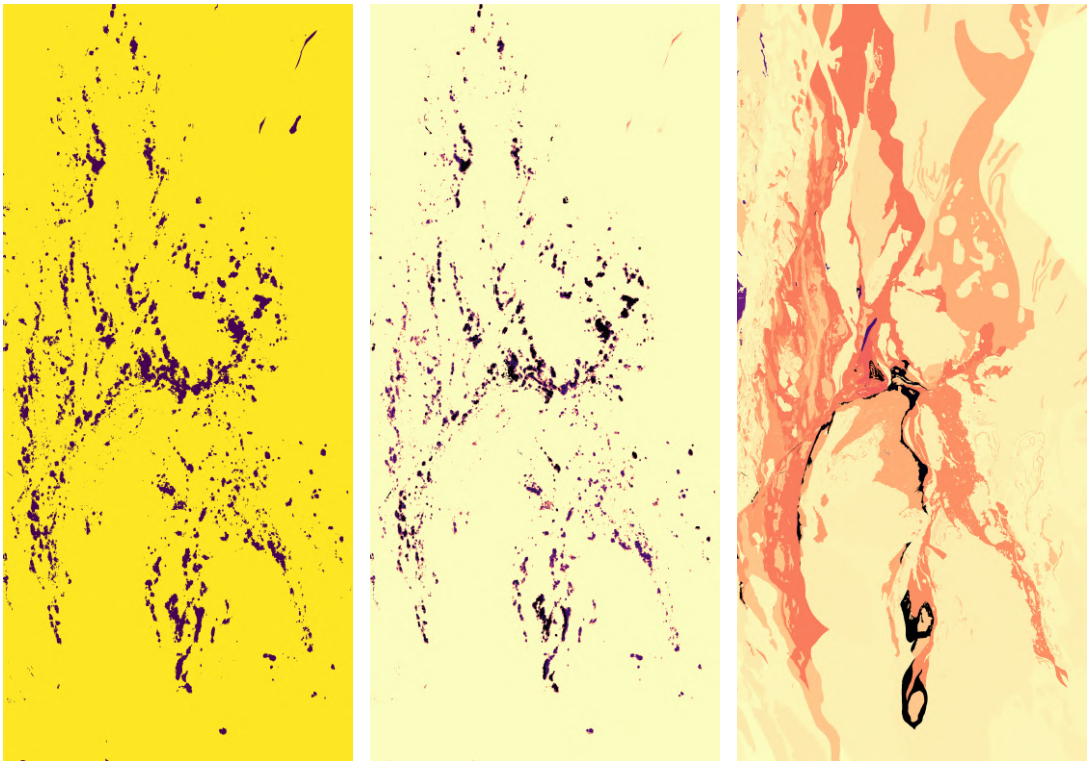


(a) Predicted anomalies.

(b) Anomalous distance.

(c) Mean anomalous distance.

Figure 8.2.: Predicted anomalies in the interpreted geological map ($nu = 0.1$). See description in Section 8.2 for details.



(a) Predicted anomalies. (b) Anomalous distance. (c) Mean anomalous distance.

Figure 8.3.: Predicted anomalies in the interpreted geological map ($nu = 0.01$). See description in Section 8.2 for details.

# 9. Modelling Geochemistry

The section describes an approach to the prediction of *in situ* soil geochemistry even under deeper cover by way of investigating the relationships between surface geochemical measurements and geophysical data.

## 9.1. Model

A random forest regressor was used to model geochemistry across the Quamby region based on subsurface covariates only. Two separate models were tested: one with magnetics and gravity covariates; and a second including the addition of wavelet transformations of these features.

Although we are interested in the Quamby area, the models were trained on surface observations across the larger Eastern Succession region believed to be of similar geology. The increase in the number of training points was necessary to improve the model performance. Figure 9.1 shows the training data used and the Quamby region for which the geochemical element concentrations were predicted is outlined.

## 9.2. Training data

The target *in situ* soil geochemistry data was obtained from the Queensland Government Department of Natural Resources and Mines [32]. The dataset contains sampled geochemistry across a range of mediums (e.g., soils, sediments, rock chips) and each with a range of sampling methods (e.g., mesh size for filtering soil samples). The dataset was separated into groups to generate subsets of consistent/comparable values. For soils, the data was grouped by '`Data Type`' and '`Mesh`' and all units were converted be consistent.

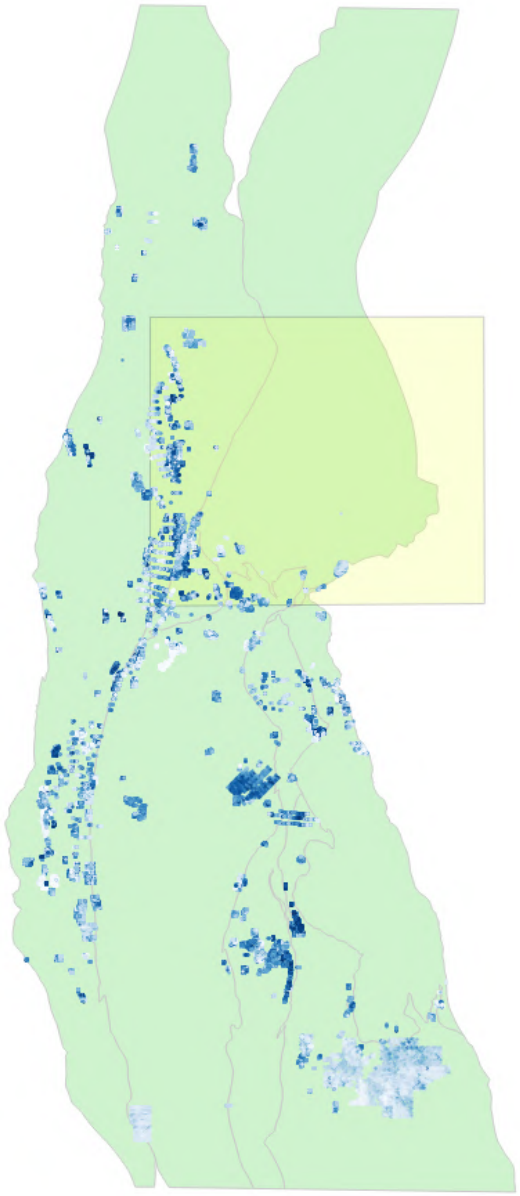


Figure 9.1.: Geochemistry model training points (showing Cu values) over the larger Eastern Succession and the Quamby region outline.

The following steps were taken to create the training dataset for each element (Cu in soil used as an example here):

1. Filter '`job_an`' where '`Column_Name == Cu`'
2. Join this to '`job`' on the '`Alias_Job`' field
3. Join this to '`s_Soil`' on the '`Job_No`' field
4. Group by '`[Data_Type, Mesh]`' and select the largest group
5. Remove points outside the training area
6. Convert values to numbers and remove any '`NaN`'s
7. Convert *below detection limit* values to half the detection limit
8. Make the units consistent (remove samples with no units)
9. Remove any entries where the '`value = 0.0`'
10. Convert from `ADG66` to `GDA94` shapefile

## 9.3. Covariate shift

This particular problem suffers from the issue of "covariate shift", whereby the dataset used to train the model (surface observations) is not representative of the dataset on which the model will be queried (the geospatial data across the Quamby area). In the case of the surface observations, they have a significant selection bias, in that they were all sampled from outcropping geologies, and that each sample was taken for reasons primarily connected to the exploration of a particular target with mostly unknown additional information.

If a query dataset varies too much from the training dataset then the model is no longer applicable, and its outputs should not be used. For this problem, we estimated the amount of covariate shift by training a classifier to discriminate between the training data and the query data. If a classifier can successfully tell the difference, we can assume the distributions are not the same and therefore the predictions made remain unreliable.

For the geochemistry datasets, a logistic regressor classifier was used to discriminate between the training and query datasets. For the *magnetics and gravity* covariates only, the classifier performed poorly, with an average accuracy across the elements of ~0.55, however with the addition of wavelet features the classifier was able to discriminate between the two distributions very well, with an accuracy of ~0.95. The addition of wavelet features makes the model more specialised for the training set data as the training samples become more distinct with respect to the query data overall.

| Element | $R^2$ score (magnetics and gravity) | $R^2$ score (magnetic and gravity + wavelets) |
|---|---|---|
| Au | 0.44 | 0.51 |
| Cu | 0.65 | 0.72 |
| Pb | 0.69 | 0.79 |
| Zn | 0.61 | 0.68 |

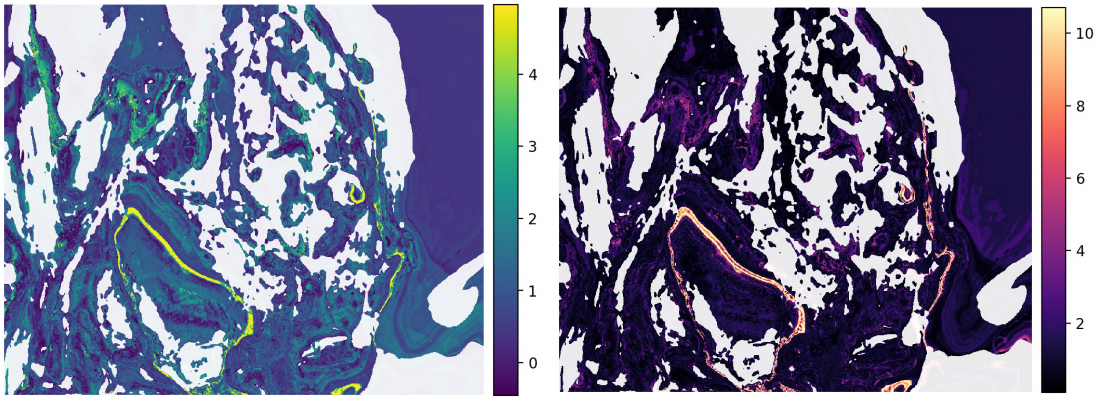
Table 9.1.: Cross-validation results for geochemistry regression models (one model using only magnetics and gravity covariates, and a second model with the addition of wavelet features).

## 9.4. Geochemistry predictions

Table 9.1 shows the cross-validation results for both models. As expected the addition of wavelets to the geophysical covariates improved the cross-validation performance of the model (e.g for Cu, $R^2$ increased from 0.65 to 0.72). However, as stated in the section above, this comes at the cost of the model becoming less general. In other words, the increase in performance is applicable only to samples which look like the samples the model was trained on and there are less of these as we include more covariates.

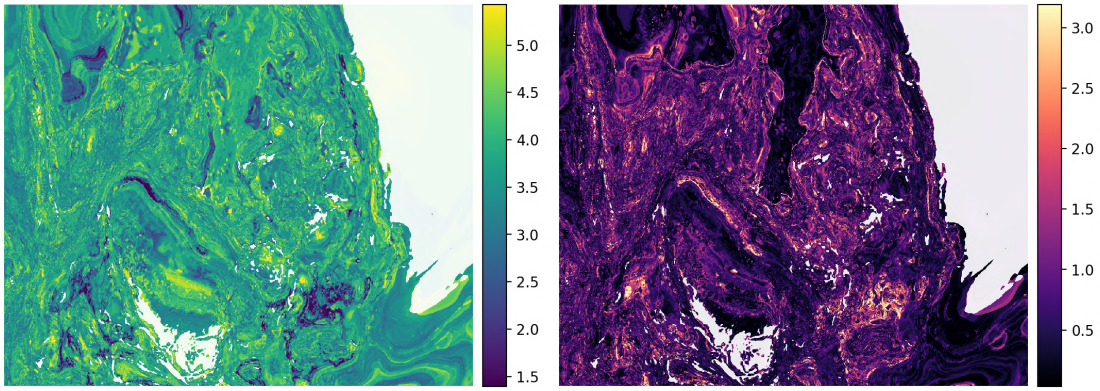
The prediction maps generated by the model are shown in Figures 9.2 and 9.9. Both the estimate (log concentration) and uncertainty (variance) produced by the models are shown for each element. The output maps are masked by the results of the covariate shift classifier (i.e. areas of the query dataset which were classified as not part of the training dataset have been masked out). This indcates where in the Quamby region the model is applicable. Even applying a generous threshold on the output of the covariate shift classifier for the *wavelets included* model, produces a very small area for which it may be considered valid.

In terms of performance, the model is more successful in modelling the concentrations of Cu and Pb. The relatively low $R^2$ scores and the additional covariate shift problem should be taken into consideration when interpreting the outputs. The performance of such models relies both on an underlying relationship between the covariate and target data and a representative training dataset. The biased nature of available datasets will have a negative impact on the ability of a model to make predictions across new areas.



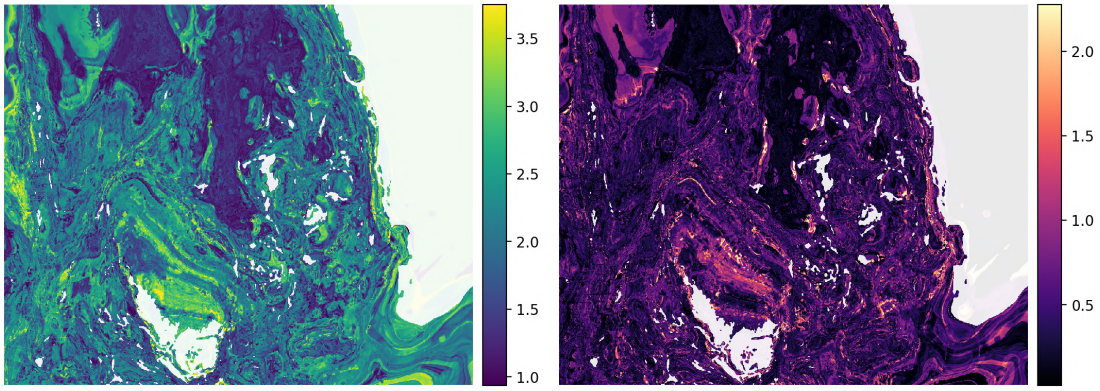

(a) log(Au) prediction. (b) log(Au) uncertainty.

Figure 9.2.: Quamby Au prediction/uncertainty (magnetics and gravity covariates, masked by covariate shift).



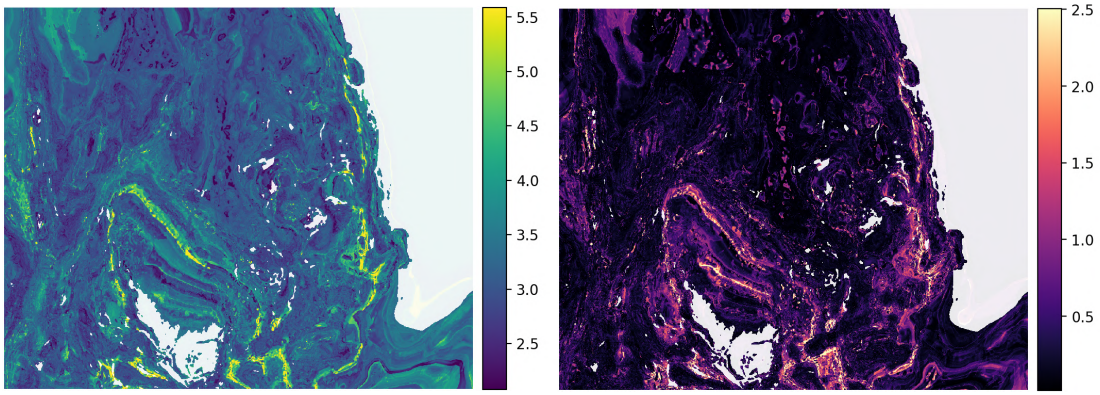

(a) log(Cu) prediction. (b) log(Cu) uncertainty.

Figure 9.3.: Quamby Cu prediction/uncertainty (magnetics and gravity covariates, masked by covariate shift).



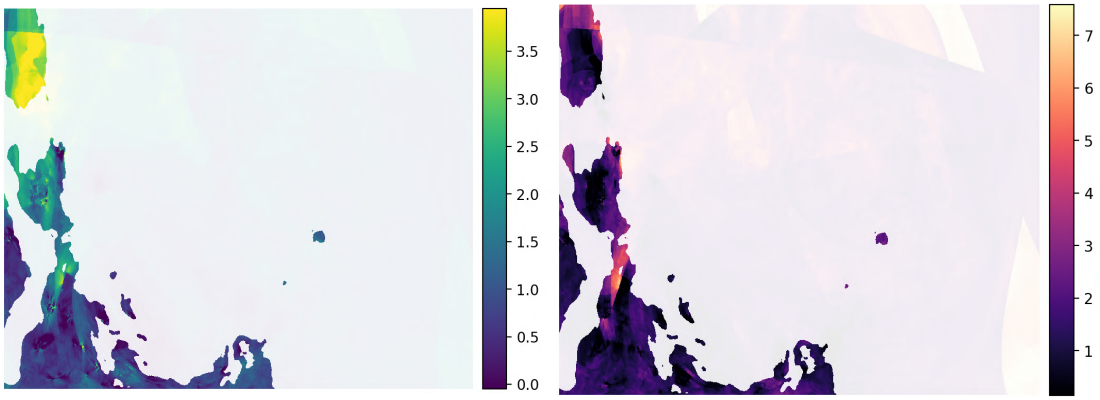

(a) log(Pb) prediction. (b) log(Pb) uncertainty.

Figure 9.4.: Quamby Pb prediction/uncertainty (magnetics and gravity covariates, masked by covariate shift).




(a) log(Zn) prediction. (b) log(Zn) uncertainty.

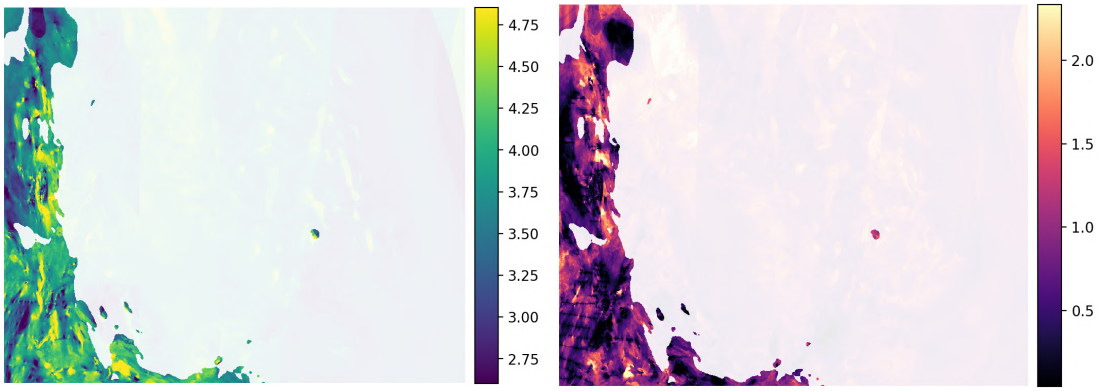
Figure 9.5.: Quamby Zn prediction/uncertainty (magnetics and gravity covariates, masked by covariate shift).




(a) log(Au) prediction.

(b) log(Au) uncertainty.

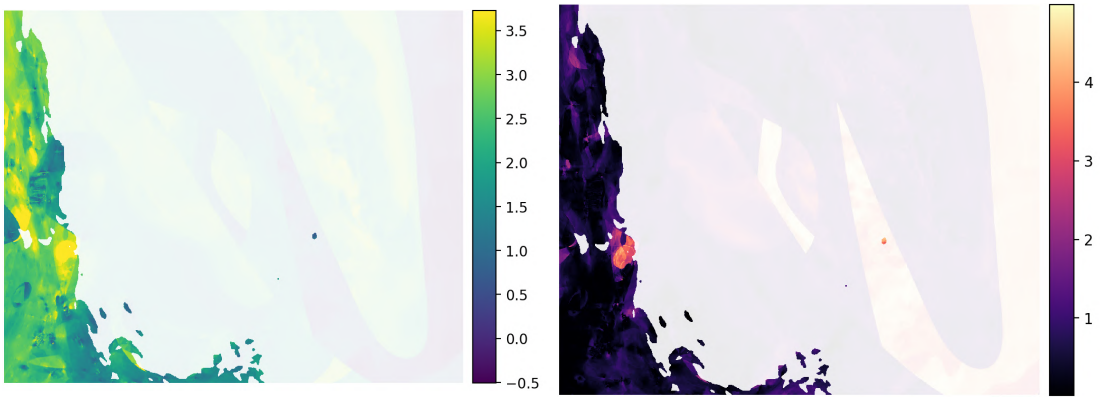
Figure 9.6.: Quamby Au prediction/uncertainty (magnetics and gravity covariates with wavelet features masked by covariate shift).




(a) log(Cu) prediction.

(b) log(Cu) uncertainty.

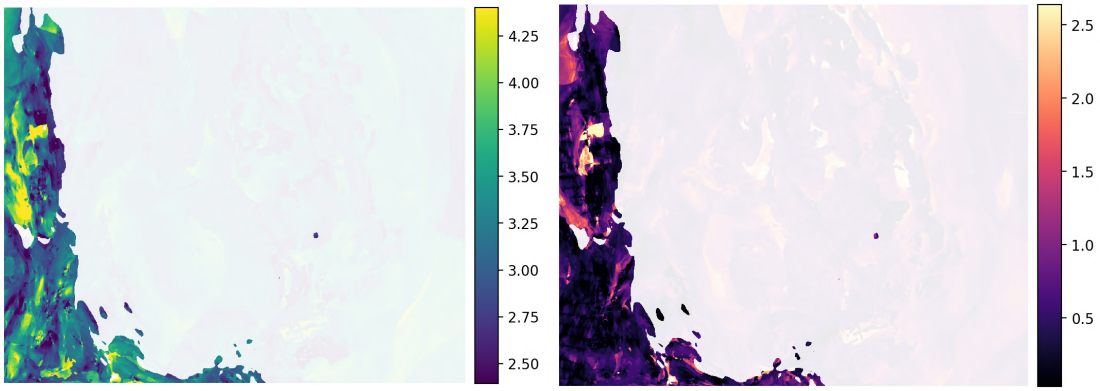
Figure 9.7.: Quamby Cu prediction/uncertainty (magnetics and gravity covariates with wavelet features masked by covariate shift).




(a) log(Pb) prediction.

(b) log(Pb) uncertainty.

Figure 9.8.: Quamby Pb prediction/uncertainty (magnetics and gravity covariates with wavelet features masked by covariate shift).




(a) log(Zn) prediction.

(b) log(Zn) uncertainty.

Figure 9.9.: Quamby Zn prediction/uncertainty (magnetics and gravity covariates with wavelet features masked by covariate shift).

## 9.5. Geological interpretation

The prediction of element concentrations in the Quamby area mainly aimed at re-evaluating the mineralisation potential of the area. As indicated by previous studies, targeting and exploration in the area heavily depend on the use of geophysical data in all of the covered areas. While geochemistry was used with success in the outcropping areas to indicate altered and mineralised rocks directly, geophysics mostly focused on the indirect mapping of structural features which were shown to control the mineralisation. The predictions containing the wavelet features provide better predictive accuracy at the expense of less available detail and the ability to predict reliably under cover. The classifications not utilising the wavelets show more detail and seem to be able to predict better undercover; however, the accuracy of predictions will be overall poorer. The following paragraphs will mostly focus on models that include all of the variables, including wavelet derived layers.

The gold concentration models provide the lowest prediction accuracies mainly due to the abundance of below detection values in the sample set as well as the natural nuggety behaviour of gold. The highest gold concentrations are predicted in Mary Kathleen Domain Capturing the Little Eva, Cabbage Tree Creek Prospect, and Blackard. Slightly lower concentrations are found further down along the Mount Rose Bee Fault. Of interest might be two anomalies found along the Quamby Fault in the south-western part of the study area. These are, however associated with relatively high uncertainty and do fall into the unreliable domain due to covariate shift.

The copper concentration model shows the second best cross-validation results. This might be the result of higher number of the training samples which is in turn caused by mostly copper focussed exploration efforts in the area. The model captures well most of the deposits and occurrences or shows elevated values immediately next to them, for example, Ernest Henry, FC4 Prospect (covariate shift indicated). The covariate shift evaluation, however, renders most of the eastern part of the study area as different from the training data and thus unreliable the only patch that stands out as not significantly different from the training data domain is around Ernest Henry. The model does not perform well in the south-eastern part of the Quamby area where neither the minor occurrences nor the larger prospects and deposits like Monakoff, Lorena are detected. This is also true of the model excluding the wavelet features and might be in general caused by the different geophysical nature of the geological domains hosting the deposits and lack of training data for these deposits/domains. The uncertainty

image shows as in the case of other elements, the dependence of the model on training sample location and density.

The best overall prediction accuracies were obtained for the lead model. The concentration map shows probable oversimplification of the model and over-dependence on the wavelet layers which can be caused by the fact that no major lead deposits were found in the area and so lead values might not be indicative anomalies. Primary lead distribution should however be best preserved out all of the elements through weathering. Some possible anomalies were identified in the western part of the area in the Mount Albert Group around the Dougald River Deposit. The central and eastern anomalies look mostly like artefacts and cannot be evaluated as real as no training data exist in the eastern part of the study region. These anomalies are correctly masked by the covariate shift estimator.

The zinc concentration map provides the third highest cross-validation score of 0.68. The map again highlights the Mount Albert Group and areas around Dougald River Deposit, as well as further locations, found south along the Mount Rose Bee Fault zone towards the Milo Beds unit in the southwest. Most of the higher concentration locations in the central and eastern part of the area also show very high uncertainties in these predictions and are again identified as regions affected by covariate shift.

# 10. Conclusions and recommendations

The main purpose of this project was to demonstrate applications where data-driven techniques have the potential to improve, aid, or automate existing processes around the preparation and validation of geological datasets and interpretations. The use of data-driven techniques across three main areas of geological investigations and characterisations were considered: prospectivity mapping, classification of lithostratigraphic units, and geochemistry predictions.

With regards to prospectivity mapping, *weights of evidence* (WoE) modelling is already a widely used data-driven approach. This method has several shortcomings that relate to the assumptions required of the training data and the inability of the model to generalise well to new data outside the training dataset. We compared this traditional approach to a classical machine learning method which does not have the same limitations as WoE and showed improved performance. Unfortunately, modelling prospectivity based on known locations of mineral deposits generally suffers from issues related to the training dataset (e.g. low number of training samples or lack of negative/unmineralised training samples), which cannot be solved entirely by using any method. Nevertheless, several novel machine learning algorithms can be considered for future applications for this type of problem.

In the area of geological mapping, we considered a number of ways in which geospatial data may help in automating the process and provide additional guidance. Separate approaches using both the existing interpreted geological map and the surface observations investigated the influence of training data distribution, choice of covariates, and model type on the quality of the predictions. While both surface observations and geophysical datasets are informative for automated geological mapping, the analyses highlighted the importance of a sufficient and representative training dataset for a model's ability to produce meaningful outputs. This means having "ground truth" observations which span both the distribution of geospatial data available and the range

of geological features of interest.

In the specific case limiting the geological classification problem to granite bodies only, the results indicate that automated detection of outcropping granites is possible given remote-sensing and geophysical data. Using classifications made from surface observations alone (collected in the process of geological mapping), reasonable *first-pass* maps of outcropping granites can be auto-generated, with larger granite bodies being particularly well identified. Furthermore, sampling information from a published geological map showed the potential of the approach given a more extensive training dataset with more evenly distributed samples.

An alternative approach to using data-driven models in the geological characterisation workflow is to assess manually produced interpretative maps for inconsistencies. An anomaly detection algorithm, applied to the existing solid geological map, that used the geophysical data as supporting information is able to identify regions which have potentially been misclassified. This output can be used by geologists to quickly highlight areas which may need reassessment or further data gathering. This experiment showed that, in addition to potentially automating tasks involving geological interpretations, data-driven approaches can also augment and/or validate existing workflows.

The final problem considered in this project was that of using geospatial datasets and surface observations for predicting geochemistry across the region. The results showed a positive correlation which captured most of the anomalies related to mineralised areas. Within the training set, the model was able to identify a relationship between the geophysical data and selected analysed elements in collected soil samples. However, the issue of covariate shift represented a significant limiting factor. The covariate shift is directly connected to the biased training set targeted towards outcropping areas and regions around the discovered mineral occurrences, geophysical anomalies potentially related to hydrothermal alteration or mafic and felsic intrusions, and prospective structural features. The clear distinction between the training and query datasets, especially for models that include additional wavelet features means that the models are not suitable for predictions across the entire Quamby area. While potentially valuable information regarding geochemistry can be obtained from the geophysics, a more uniformly distributed geochemical training dataset would significantly enhance the prediction. In essence, a purely data-driven approach has limited capability to model relationships within regions of the feature space where it does not have training points to learn from.

The problem of sample bias was common to all experiments. In all areas, algorithms

performed better where it was possible to sample across the region and/or feature space uniformly. As shown in the geochemical prediction example, data-driven techniques can be used to identify areas where training data is insufficient. Even though it must be acknowledged that obtaining relevant geochemical samples from areas of deep cover is a significant obstacle new research targeting chemical and physical paleo interfaces in these cover sequences could provide much needed additional information. A future area of consideration should to use machine learning techniques during the design and execution of data-gathering and survey processes in order to reduce sample bias. Such an "active sampling" approach can be guided by data-driven models which include prediction uncertainties in order to feed back data which will improve both the model performance and geological understanding of a region. Experiments such as these can also highlight the usefulness of various geospatial covariate datasets and improve future data collection practices.

The availability of large datasets that map chemical and physical properties of rocks lends itself to the use of this valuable information in most aspects of geological mapping and mineral exploration. For government entities like geological surveys, it is crucial to both store and serve such data to the public. Furthermore, this data needs to be employed and integrated into mapping programs, and the design of new data acquisitions. It is therefore vital to develop and acquire tools that facilitate the simultaneous analysis of multi-source data. In practice, this means taking into consideration the FAIR data principles [1], collecting machine-readable metadata which minimises the need for human intervention, and supporting access from computing resources capable of running data-driven models on large datasets.

Overall, the project has demonstrated the potential of data-driven modelling applied to large and multi-source geological datasets. While certainly not exhaustive, we have highlighted many areas where machine learning models can be used to augment, automate, or validate existing processes around the geological interpretation of these datasets. Although this work was limited to the NWMP region in Queensland, similar investigations could be conducted in other regions, and data processing and modelling techniques would benefit from cross-region comparisons. Future work in this area should identify and develop broadly applicable and repeatable processes which maximise geological information from existing datasets and inform decisions about further data collection.

[1] Findability, accessibility, interoperability, and reusability [33]

# Appendices

# A. Textural features

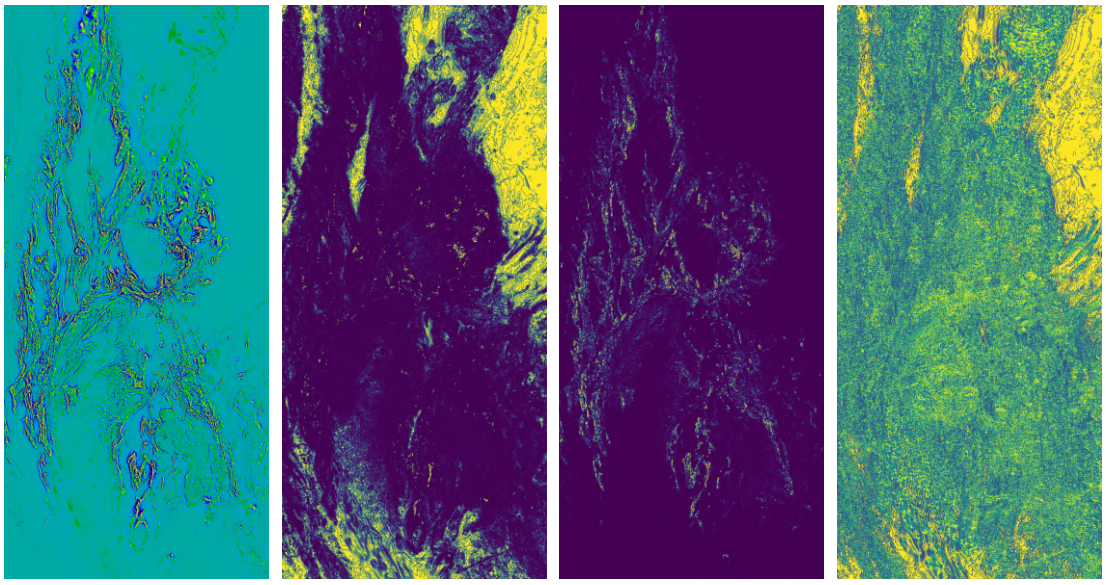
In addition to wavelet features (Section 4.3), textural analysis can also be used to capture spatially varying properties local to a point. The inclusion of textural features derived from the original geospatial feature data in a data-driven model allows the algorithm to account in some way for the spatial context surrounding an individual data point.

The most texturally relevant dataset for the problem of identifying distinct geological regions was deemed to be the first vertical derivative (1VD) magnetics dataset (Figure 4.1c). A visual assessment of "texture" within the magnetics dataset is already part of the manual process of geological classification traditionally performed by the geologist.

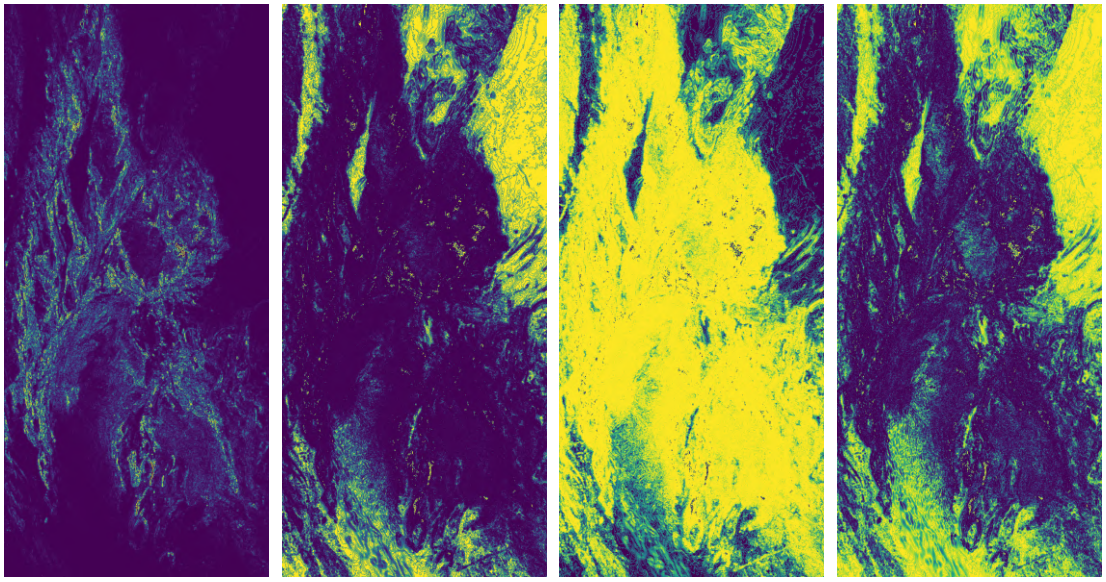
Grey-level Co-occurrence Matrix (GLCM) features (or Haralick textures) are commonly used in geospatial and remote sensing image processing applications to quantify a range of local spatial properties [34, and references within]. A number of GLCM based features were calculated for the 1VD magnetics dataset across various scales. The properties used to derive these features are shown in Table A.1. The GLCM processing resulted in 168 different features. As an example, the seven features derived for a window size of three by three pixels, a distance of one and symmetric angle of zero (north-south) are shown in Figure A.1.

To remove inherent redundancy and reduce the number of covariates provided to the model Principal Component Analysis (PCA) was used to reduce the number of inputs from the 168 GLCM features down to the 12 features shown in Figure A.2.

While the inclusion of the GLCM features here generally improved the cross-validation performance of the model, further investigation could be conducted to determine which textural properties are most useful for identifying particular aspects of geology. An alternate approach to using prescribed textural measures such as GLCM as features could be to use a deep learning model. In such a model, spatial features can be learnt as part of the model training process; however, they do require large training data sets.

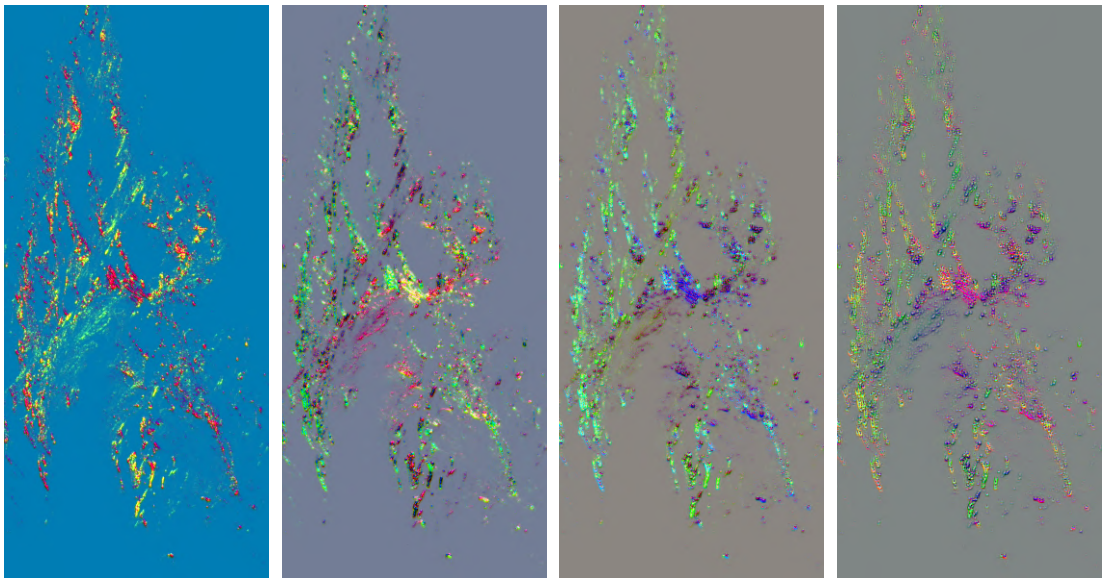


(a) Mag 1VD 512 digitised (b) ASM (c) Contrast (d) Correlation



(e) Dissimilarity (f) Energy (g) Entropy (h) Homogeneity

Figure A.1.: Example GLCM textures derived from a 512 bin digitisation of the magnetics 1VD data. The GLCM was created with a windows size of 3 pixels, a distance of 1 pixel and symmetric angle of 0 (north-south).



(a) Dimensions 1–3. (b) Dimensions 4–6. (c) Dimensions 7–9. (d) Dimensions 10–12.

Figure A.2.: The first 12 PCA dimensions of the set of Haralick textures calculated from the 1VD magnetics data. Three dimensions are shown at a time as RGB images.

Table A.1.: GLCM feature parameters.

| **Property** | **Values** |
|---|---|
| angles | $0, \pi/4, \pi/2, 3\pi/2$ (symmetric) |
| distances | 1 |
| window sizes | $3, 5, 9, 11, 17, 21$ |
| levels | 512 |
| textures | ASM, Contrast, Correlation, Dissimilarity, Energy, Entropy, Homogeneity |

# References


[1] C.T. McCuaig, S. Beresford, and J. Hronsky. "Translating the mineral systems approach into an effective exploration targeting system". In: *Ore Geology Reviews* 38.3 (2010), pp. 128–138.

[2] S.A. Drury. *Image interpretation in geology*. Nelson Thornes UK, 2001.

[3] F.D. Van der Meer et al. "Multi-and hyperspectral geologic remote sensing: A review". In: *International Journal of Applied Earth Observation and Geoinformation* 14.1 (2012), pp. 112–128.

[4] H. Saadat et al. "Landform classification from a digital elevation model and satellite imagery". In: *Geomorphology* 100.3-4 (2008), pp. 453–464.

[5] P. Betts et al. "Kinematic analysis of aeromagnetic data: Looking at geophysical data in a structural context". In: 11.4 (2007), pp. 582–583.

[6] V. Metelka et al. "A geophysically constrained litho-structural analysis of the Eburnean greenstone belts and associated granitoid domains, Burkina Faso, West Africa". In: *Precambrian Research* 190.1-4 (2011), pp. 48–69.

[7] G.P.T. Spampinato et al. "Structural architecture of the southern Mount Isa terrane in Queensland inferred from magnetic and gravity data". In: *Precambrian Research* 269 (2015), pp. 261–280. ISSN: 03019268. DOI: 10.1016/j.precamres.2015.08.017.

[8] P. An, C.F. Chung, and A.N. Rencz. "Digital lithology mapping from airborne geophysical and remote sensing data in the Melville Peninsula, Northern Canada, using a neural network approach". In: *Remote Sensing of Environment* 53.2 (1995), pp. 76–84.

[9] V. Rodriguez-Galiano et al. "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines". In: *Ore Geology Reviews* 71 (2015), pp. 804–818.

[10] C. Kirkwood et al. "A machine learning approach to geochemical mapping". In: *Journal of Geochemical Exploration* 167 (2016), pp. 49–61.

[11] M.J. Rubenach et al. "Age constraints on the tectonothermal evolution of the Selwyn Zone, Eastern fold belt, Mount Isa Inlier". In: *Precambrian Research* 163.1-2 (2008), pp. 81–107.

[12] Geological Survey of Queensland. *North-West Queensland Mineral and Energy Province Report*. Queensland Department of Employment, Economic Development and Innovation, Brisbane. 2011.

[13] M.A. Etheridge, R.W.R. Rutland, and L.A.I. Wyborn. "Orogenesis and tectonic process in the Early to Middle Proterozoic of northern Australia". In: *Proterozic Lithospheric Evolution* 17 (1987), pp. 131–147.

[14] D.R.W. Foster and J.R. Austin. "The 1800-1610 Ma stratigraphic and magmatic history of the Eastern Succession, Mount Isa Inlier, and correlations with adjacent Paleoproterozoic terranes". In: *Precambrian Research* 163.1-2 (May 2008), pp. 7–30.

[15] M.L. Greenwood and C.R. Dhnaram. *3D mineral potential of the Quamby area*. Queensland Minerals and Energy Review Series. 2013.

[16] T.J. Beardsmore, S.P. Newbery, and W.P. Laing. "The Maronan Supergroup: an inferred early volcanosedimentary rift sequence in the Mount Isa Inlier, and its implications for ensialic rifting in the Middle Proterozoic of northwest Queensland". In: *Precambrian Research* 40-41.C (1988), pp. 487–507.

[17] D.H. Blake. *Geology of the Mount Isa inlier and environs, Queensland and Northern Territory*. Vol. 225. Australian Govt. Pub. Service, 1987.

[18] C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[19] R. Cant. *Queensland Gravity Grid*. accessed 9-8-2018. 2014.

[20] M. Greenwood. *Rad_Dose, Rad_Pot_pct, Rad_Tho_ppm, Rad_Ura_ppm*. accessed 9-8-2018. 2018.

[21] M. Greenwood. *Queensland Merged RTP, Queensland Merged Magnetic 1VD*. accessed 9-8-2018. 2018.

[22] T.J. Cudahy et al. *3D Mineral mapping of Queensland - Version 2 ASTER and related geoscience products. v1.* https://data.csiro.au/dap/landingpage?pid=csiro:20912. 2017.

[23] J.C. Gallant et al. *1 second SRTM Derived Digital Elevation Models User Guide.* 2011.

[24] J.C. Gallant and T.I. Dowling. "A multiresolution index of valley bottom flatness for mapping depositional areas". In: *Water resources research* 39.12 (2003).

[25] B. McCormack. *Queensland geological mapping digital data - November 2012.* accessed 9-8-2018. 2017.

[26] M. Greenwood. *GSQ Site Observation data.* accessed 9-8-2018. 2017.

[27] J. Tang. *Mount Isa East 2016 Data Package.* accessed 9-8-2018. 2017.

[28] GDAL/OGR contributors. *GDAL/OGR Geospatial Data Abstraction software Library.* Open Source Geospatial Foundation. 2019. URL: https://gdal.org/gdaldem.html.

[29] A. Burkov. *The Hundred-Page Machine Learning Book.* 1st ed. Kindle Direct Publishing, 2019. ISBN: 9781790485000.

[30] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python ". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: https://scikit-learn.org/stable/documentation.html.

[31] D. Zhou et al. "Learning with local and global consistency". In: *Advances in Neural Information Processing Systems 16.* MIT Press, 2004, pp. 321–328.

[32] Department of Natural Resources Queensland Government and Mines. *QDEX Data.* URL: http://qdexdata.dnrm.qld.gov.au (visited on 08/14/2018).

[33] Australian National Data Service (ANDS). *The FAIR data principles.* 2017. URL: https://www.ands.org.au/working-with-data/fairdata (visited on 06/18/2019).

[34] M. Hall-Beyer. *GLCM texture: a tutorial.* Department of Geography, University of Calgary, Mar. 2017. URL: https://www.researchgate.net/publication/246325910_GLCM_texture_a_tutorial.

# Disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

CONTACT US

**t** 1300 363 400
**e** enquiries@csiro.au
**w** www.csiro.au

YOUR CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.

FOR FURTHER INFORMATION

**CSIRO Minerals Resources Flagship**
Andrew Rodger
**t** +61 8 6436 8512
**e** andrew.rodger@csiro.au
**w** Minerals Resources Flagship