Lithological identification and coal quality prediction through machine learning methods using geophysical log data

Arek Chouzadjian and Kane Maxwell

2022-10-27

Abstract

Mining is a crucial global industry, and a major part of that industry's exploration activities is rock type (lithological) classification and mineral quality analysis. However, these activities are typically done manually, which is time-consuming and prone to inaccuracies. The aim of this report is to show that, in the context of the coal industry, these activities can be automated using machine learning methods that take advantage of open-source geophysical log data to predict lithologies and mineral quality, and that these predictions can be displayed in a prototype dashboard for industry use. The analysis found that supervised learning methods such as random forests, gradient-boosted trees and support vector machines all display quite impressive performance for lithological classification, especially for coal, where as unsupervised learning methods such as k-means clustering perform poorly. The analysis also found that multivariate and compositional methods using random forest and principal components regression are highly accurate at predicting coal quality measures. The results suggest that these methods are viable within the industry, allowing companies to save time and money. Additionally, these methods can be adapted to other mineral varieties, to other activities such as mapping and safety analysis, and can also be extended to non-mining geo-investigative work.

Background and Motivation

Lithological classification

Mining is a critical global industry, extracting minerals from the ground that are required for a wide array of industrial processes in modern society. Mineral exploration is an essential aspect of this industry, and consists of a range of activities to determine whether there are minerals of interest in the ground that can be mined (Haldar, 2013). These activities include mapping, surveying of the ground (either surface-based or airborne surveying), testing of water and soil samples, and drilling (Haldar, 2013). Drilling in particular is important in later, targeted stages of mineral exploration. This is because it allows geologists to determine whether the mineral of interest is present, how much is present, how it is distributed, as well as the quality of the mineral and whether it is appropriate for economic extraction (Firth, 1999).

Drilling is the process of penetrating through the ground and extracting rock samples from various depths beneath the surface for confirming the geology beneath (Firth, 1999). Conducted alongside drilling is what is known as "well logging", which is the process of collecting detailed information on the geological formations penetrated by drilling. This involves recording geophysical log data on rock types (lithologies) and mineral quality through the use of specialised equipment in the boreholes created by drilling. This equipment collects data relating to gamma ray levels, density, resistivity, sonic readings, and other measures (Firth, 1999). Different rock types and minerals often have distinctive geophysical signatures; this data, along with the physical samples from drilling, are often used by geologists to assist in determining rock types, their distribution/depth, and their quality. Conventionally, this is done manually by geologists, who bring to bear their own experience and expertise in the ultimate classification decision (Rau et al., 2022).

While geophysical log data and physical samples are indispensable in evaluating complex lithologies (Zhong

et al., 2020), manual interpretation of this data is time-consuming and tedious (Kumar et al., 2022). This is due to the fact that geophysical log data behaves differently depending on the depth of and conditions within a borehole (of which there are many variations), local ground conditions, and the associated fact that different rock types are not always easily distinguishable on the basis of the data (Firth, 1999). Furthermore, manual interpretation is prone to uncertainty and inaccuracies, as different geologists may have different interpretations of geophysical log data, depending on their experience (Rau et al., 2022). Finally, drilling is a highly expensive and intrusive exercise, which mining companies prefer to employ selectively so as to minimise costs, and is usually only conducted after other preliminary assessments have been carried out.

In response to these issues, in recent decades there has been a great deal of interest in the automation of these classification and quality assessment processes, as well as extensions to activities such as lithological mapping. Mining companies collect a huge amount of data, much of which is not used or harnessed to its full potential. This data, such as geophysical log data, can be exploited through the use of various machine learning methods (Bhattacharya et al., 2016), with the goal of making lithological classification faster and more accurate. These methods also have the potential to reduce costs for resource firms by allowing them to minimise drilling, or engage in types of drilling that are less financially onerous (Zhou et al., 2020). This is especially true when extensions to 3D-visualisation and mapping are considered (Zhou et al., 2020). Accordingly, there has been an explosion in research into the use of machine learning in exploration activities, as well as adoption by resource firms, especially in the petrochemical (i.e. oil and gas) sector.

The research base into the use of machine learning algorithms for lithological classification is diverse and well-established. Most studies involve the use of geophysical log data collected from a number of boreholes located in a specific exploration area. Research has focused on a wide range of mineral targets, including oil, gas, coal, uranium, and others (Maxwell et al., 2019; Bressan et al., 2020; Sun et al., 2022). All manner of machine learning classifiers have been trained and tested on geophysical log data; both supervised and unsupervised learning methods have been investigated. Within the body of research into the use of machine learning for lithological classification, there is no one method that is more popular than any other; given the amount of research into the area, this is not necessarily surprising. The accuracy of the classifiers appears to be most linked to a combination of the lithologies present, local conditions, and data availability.

Specific learning methods that have been commonly used in research include supervised learning methods such as random forests (RF), gradient-boosted trees (GBT), neural networks (NN) and support vector machines (SVM), as well as unsupervised learning methods such as self-organising maps (SOM) and multi-resolution graph-based clustering (MRGC). RF and GBT have demonstrated particularly strong results across a variety of contexts (Maxwell et al., 2019). In a study investigating lithological classification in gas fields in China, RF and GBT outperformed NN and SVM (Xie et al., 2018). In the same study, GBT was the best of all methods investigated in classifying sandstone varieties. In another study examining lithological classification using data from offshore wells, RF again outperformed NN and SVM (Bressan et al., 2020). RF also outperformed SVM in accuracy, 79% to 74%, in a carbonate sediment classification task that employed well log data (Insua et al., 2015).

In terms of the other learning methods used, SVM have been a very popular choice for researchers in investigating the use of machine learning methods in lithological classification. The success of SVM in classifying rock types, as opposed to other learning methods, has been mixed. In a study looking at lithological classification in a sandstone reservoir, SVM outperformed NN (Al-Anazi et al., 2010). In Kumar et al. (2022), SVM, NN, RF and GBT were trained on well logs for the purpose of predicting coal and non-coal rock types in the Talcher coalfield in India; all four methods exhibited similar classification performance. In the classification of mudstone types, SVM outperformed NN and unsupervised learning methods like SOM and MRGC (Bhattacharya et al., 2016). However, in Bressan et al. (2020), SVM performed particularly poorly compared to even NN.

Overall, out of the various machine learning methods used in this area of research, RF, GBT and SVM display the best combination of breadth of application/investigation, as well as performance. These three methods form a strong foundation upon which to base multi-class rock type classification, and subsequently form the basis of the modelling techniques used for such classification in this report.

Coal quality prediction

A key aspect of assessing whether a mineral is appropriate for economic extraction is examining its quality. In the context of this report, coal is the relevant mineral of interest. Traditionally, coal quality parameters such as ash content, relative density, volatile matter, fixed carbon and moisture content are derived from measurements made on physical core samples (Firth, 1999). Many geophysical log measurements are sensitive to elements of coal that control these traditional quality parameters (Firth, 1999). The subsequent idea is that geophysical log data could be used to predict coal quality measurements, and that this prediction could be done using various machine learning regression methods. This approach would have obvious benefits, in terms of not having to rely on the treatment of physical core samples, as well as automating and improving the quality assessment process. This idea is also applicable to other types of minerals.

Typically, coal quality modelling using physical core sample data has been conducted through the use of various weighted-average models, such as inverse distance weighting, and various types of kriging (Jeuken et al., 2020). Geophysical log data have been used to supplement these models, in some cases enabling improvements to coal quality parameter estimation (Jeuken et al., 2020). Alternatively, models have been built for coal and other mineral quality parameter estimation that use just geophysical log data, relying on relationships between the log variables and coal quality parameters for prediction. These have ranged from linear regression models (Zhou et al., 2020), to non-parametric methods such as random forests (Gordon et al., 2022), neural networks (Siregar et al., 2017, Ghosh et al., 2017) and other types of algorithms (Fullagar et al., 1999). In this report, the focus of the investigation into using geophysical log data for coal quality prediction will be centred on the latter approach; that is, modelling using just the geophysical log data for estimation.

It is important to note that, with regards to coal quality parameters, ash, volatile matter, fixed carbon and moisture content are composite in nature. In other words, these four coal quality measures add up to 100%. Most literature focuses on individual parameter modelling; the success of the modelling techniques used in existing research varies between the different quality measures. For all of the parameters mentioned, it was found that density logs are the most important for estimation, but that estimation performance can be improved by using gamma logs, as well as varying resolutions of density logs (e.g. long- and short-spaced density logs) (Zhou & O'Brien, 2016). Zhou & O'Brien (2016) found that these geophysical log variables are good predictors for ash and fixed carbon content, but they are less successful at predicting volatile matter and moisture content. In particular, these geophysical log variables are especially poor at estimating moisture content (Ghosh et al., 2017), but model performance can be improved by including other log variables such as neutron porosity (Jie et al., 2014).

In this report, it was decided to use modelling methods typically utilised in the literature to estimate the coal quality parameters, with a focus on using density and gamma logs as predictors. Specifically, RF and principal components regression (PCR) methods were used. However, whilst individual parameter prediction was undertaken using random forest models for each parameter, the main focus was on multivariate and compositional modelling methods using random forest and principal components regression, due to the compositional nature of the coal quality parameter data.

Objectives and Significance

This report has three key objectives. The first objective is to use geophysical log data, particularly focusing on the use of gamma readings, as well as short and long-density measurements, to predict lithology types. This prediction strategy is specifically directed at a complex, multi-class classification problem involving numerous different rock types. Both supervised and unsupervised machine learning methods are used to predict the lithologies, and these different methods are assessed and compared to one another in terms of performance. The second objective of the report is to use the same geophysical log data used for the lithological classification to predict coal quality measurements such as ash content, moisture content, fixed carbon content and the like. Prediction will be undertaken using different varieties of regression methods, with a focus on multivariate and compositional analysis. These methods will also be assessed and compared to one another in terms of performance. Finally, the third objective of this report is to create a dashboard that

displays the predictions and performance metrics of the different models for the lithological classification and coal quality prediction problems, where the user is able to select between different models. The dashboard will serve as a prototype for further development for potential commercial use.

The data used for the analysis is open-source, having been released in 2021 by the Geological Survey of Queensland (GSQ). This data includes information on borehole location and associated downhole geology data, incorporating lithological classification data that has been entered in manually by geologists. The data also includes geophysical log data in the form of LAS files, as well as coal quality analysis data that is derived from the manual treatment of core samples. Critically, the data used all come from boreholes on the same mine site. Further discussion of the data is contained in the data section of this report. In terms of the machine learning methods used, for the classification section, random forest (RF), gradient-boosted tree (GBT) and support vector machine (SVM) were used as supervised methods, and k-means clustering was used as an unsupervised method. These methods were chosen on the basis of the strength of research into the suitability of these classifiers for this particular problem. For the coal quality prediction section, multivariate random forest, compositional random forest, and multivariate principal components regression were used, as was individual random forest regression on each coal quality parameter by way of comparison. The multivariate/compositional methods were chosen due to their suitability to the response data, as well as the predictors.

The significance of this report is clear; using geophysical log data to predict lithologies and coal quality through machine learning methods, rather than tedious, time-consuming and inconsistent manual avenues, can make these essential elements of mineral exploration much faster and more accurate, saving mining companies time and money. Incorporating extensions to mapping and other tasks, these strategies can also benefit mining firms by allowing them to conduct less invasive mineral exploration activities. The methods seen here can also be extended to other mineral contexts, or even non-mining geological activities. In terms of the novelty of the analysis, this analysis does the following things differently to other reports in this field. First, most reports focus on simpler classification problems; either a binary (e.g. coal/non-coal) problem, or focusing on one rock type in particular, and trying to predict different gradations of it. This report seeks to classify numerous different lithologies, which is a more complex task. Second, in the literature regarding coal quality analysis, it is not common to see analyses that use multivariate and/or compositional methods to predict coal quality measurements, even though these measurements are compositional in nature. This report seeks to heavily employ multivariate and compositional analysis, which is relatively novel. Lastly, the creation of a dashboard as a prototype for further development for commercial use is also not common in the literature. Overall, this report makes an important contribution to the body of research into this area.

Data

All data related work such as data cleaning, exploratory data analysis and modelling, was undertaken using the R programming language (R Core Team, 2022). Data cleaning was conducted predominantly through the use of the tidyverse package (Wickham et al., 2022).

Data source and description

Mining companies that engage in mineral exploration are typically required to provide a range of reports to relevant government bodies in the jurisdictions in which exploration is undertaken. These bodies include geological survey organisations, which collect geoscience data and information, and the reports that mining companies provide include various types of geophysical and geochemical data. Normally, these reports are not released to the public, as they are kept confidential for commercial reasons. However, in 2020, the Queensland government changed the way confidentiality applies to these reports; subsequently, the Geological Survey of Queensland (GSQ) was able to release them the following year. Among these reports are Mining Development Lease (MDL) annual reports, which include much of the data required for this analysis, such as borehole location and geology data, geophysical log data, and coal quality analysis data. The data from one of these MDL annual reports forms the basis for the analysis in this report.

The MDL annual report selected for use in this report relates to exploration activities undertaken at the

Millennium and Mavis Downs metallurgical coal mine, 139km SW of Mackay, Queensland. The data attached to this report was extracted from GSQ's CKAN database using the ckanr (Chamberlain et al., 2021) package, and consists of predominantly text files and LAS files. The text files relevant to this report include lithological classification data; that is, for observations from different boreholes at different depths, the rock type for each observation has been recorded manually. The relevant text files also include raw coal quality data, where coal quality parameters (e.g. % ash content, % moisture content, etc...) for different samples have been recorded after physical testing, categorised by borehole and seam. Finally, coal seam picks data was also present, which details the depths at which different seams occurred in different boreholes; this dataset is important for correctly matching coal quality observations to corresponding geophysical log data.

The LAS files attached to the chosen MDL annual report contain the geophysical log data, such as gamma ray, resistivity and density measurements, which are used as the basis for the predictions in this report. "LAS" stands for Log ASCII Standard, and is the standard file format used in industry for storing geophysical log data collected from boreholes. They are distinct from "lidar" LAS files, which store remote-sensed geospatial data (Dick & Maxwell, 2022). Each borehole has its own corresponding LAS file, and within each LAS file, geophysical log measurements were recorded for different depths. These files (423 in total) were loaded into R using the lastools package (Dick & Maxwell, 2022), which is specifically designed to load LAS files into R as data tables, as well as perform other functions. Lastly, as all of these files have the exact same columns, the 423 data tables that were loaded into R were all combined together into a single dataframe of approximately 1.8 million observations.

Data cleaning

The data cleaning required to enable the separate datasets to be joined, so as to enable exploratory data analysis and predictive modelling, was quite extensive. A significant issue with the data attached to MDL annual reports is that columns referring to the same variable (e.g. borehole name, seam, etc...) are often named differently between datasets. Furthermore, levels of the same variable are also often labelled differently between datasets. There was a need to make the different datasets consistent with one another; to this end, the datasets were cleaned to make them conform with the Australian Coal Logging Standard v3.1 (Larkin & Green, 2021), and level labelling was also standardised. This enabled the different datasets to be joined effectively so that the analysis could proceed. Additionally, for applicable datasets, variables irrelevant to the analysis were dropped.

Some of the datasets required specific data cleaning. For the coal quality data, it is essential that the variables ash content, volatile matter content, moisture content and fixed carbon content, being composite percentages, add up to 100. Any observations for which these variables did not add up to 100 were dropped. Moreover, seams that had less than 30 observations were also dropped, and any rows with missing values were excluded. For the geophysical log data, nameless columns had to be removed; this sometimes happens with LAS files due to import errors. Furthermore, geophysical log variables should realistically only occur within a defined range of values, with the specific range depending on the variable. Errors in equipment readings or encoding of the data mean that some observations can fall outside of this range. Rows with observations that fell outside of the relevant defined range were dropped from the dataset.

After this process was completed the different datasets were joined to enable exploratory data analysis and predictive modelling. For the lithological classification task, the lithology (rock type) data was joined to the geophysical log data. For the coal quality prediction task, the coal quality data was joined to the seams picks and geophysical log data. Lastly, for both tasks, the size of the datasets was reduced by aggregating the geophysical log data for the groupings relevant to the analysis, and calculating the mean, first quartile and third quartile for each geophysical log variable; predictions were made using the mean values. It is important to note that one of the main limitations of the data is that not all datasets contained information on the same boreholes; subsequently, when the datasets were joined there was a significant degree of data loss, especially for the coal quality prediction task. However, this did not prevent the analysis from proceeding. It should also be noted that whilst manually-inputted lithological classification data can suffer from inconsistencies, the data still provides a strong basis for training the classifiers used in this report.

Exploratory Data Analysis

Lithological classification

The exploratory data analysis (EDA) carried out on the lithological classification dataset was quite involved. The main goals were to check for and remove outliers caused by faulty readings due to borehole/ground conditions, equipment error and inconsistent classification, determine which variables are important for prediction by looking at completeness and association with the variable, and finally examine how smaller lithology groups can be appropriately grouped together so as to create stable classes for the modelling process. All of the analysis carried out here is intended to ready the dataset for modelling, and to inform the modelling process undertaken.

The EDA for the lithological classification dataset proceeded as follows. Firstly, descriptive statistics for the mean geophysical log variables were computed. These summary statistics can be found in Table 1. The descriptive statistics featured include the minimum value of the variables, the maximum value, the mean, standard deviation and kurtosis of the variables, and finally the number of non-missing observations.

Table 1: Descriptive statistics for geophysical log variables

Variable	min	max	mean	sd	kurtosis	obs
Caliper density (mm)	855.60	1469.74	1172.19	261.89	-2.11	7
Short-spaced density (g/cm^3)	0.94	2.86	1.99	0.47	-1.40	4795
Long-spaced density (g/cm^3)	0.99	2.69	1.99	0.44	-1.43	4801
Resistivity FE1 (ohm-m)	9.51	3687.17	294.87	730.81	6.20	1401
Resistivity FE2 (ohm-m)	0.35	2857.10	769.03	1150.57	-0.69	506
Gamma radiation (API)	7.10	476.03	103.36	55.08	6.44	4857
Velocity (m/s)	114.02	7736.80	2802.80	847.20	5.48	920

For this report, the most important statistics are kurtosis, and the number of non-missing observations. Kurtosis measures the combined weight of the tails of a distribution relative to its centre. In this way, kurtosis can be used as an indicator of the presence of outliers; a high kurtosis value is indicative of outliers. The number of observations serves as a measure of the completeness of the variables. The summary statistics suggest that whilst gamma and density logs are very well represented, the other log variables are not; their observations are largely missing. This means it may be difficult to keep these variables for the modelling stage, especially if they lack association with the response. Furthermore, the kurtosis values indicate that there is a particularly high number of outliers for the gamma log variable; the reason for this requires further investigation.

The next step was the creation of boxplots for the gamma log variable and density log variables, by each lithology type. This enables the identification of the lithology types in which the outliers for gamma are occurring, potentially revealing information about their cause. However, these plots also tell us important information about the rock types themselves: which rock types are well represented in the data set, which are poorly represented, as well as which rock types exhibit similar gamma and/or density readings and can be potentially grouped together. Critically, they also provide an indication of whether the gamma and density log data have any association with the lithology types. These boxplots can be seen in Figure 1 and Figure 2. It should also be noted that as the short- and long-density logs are very highly correlated, they can be used interchangeably for visualisations.

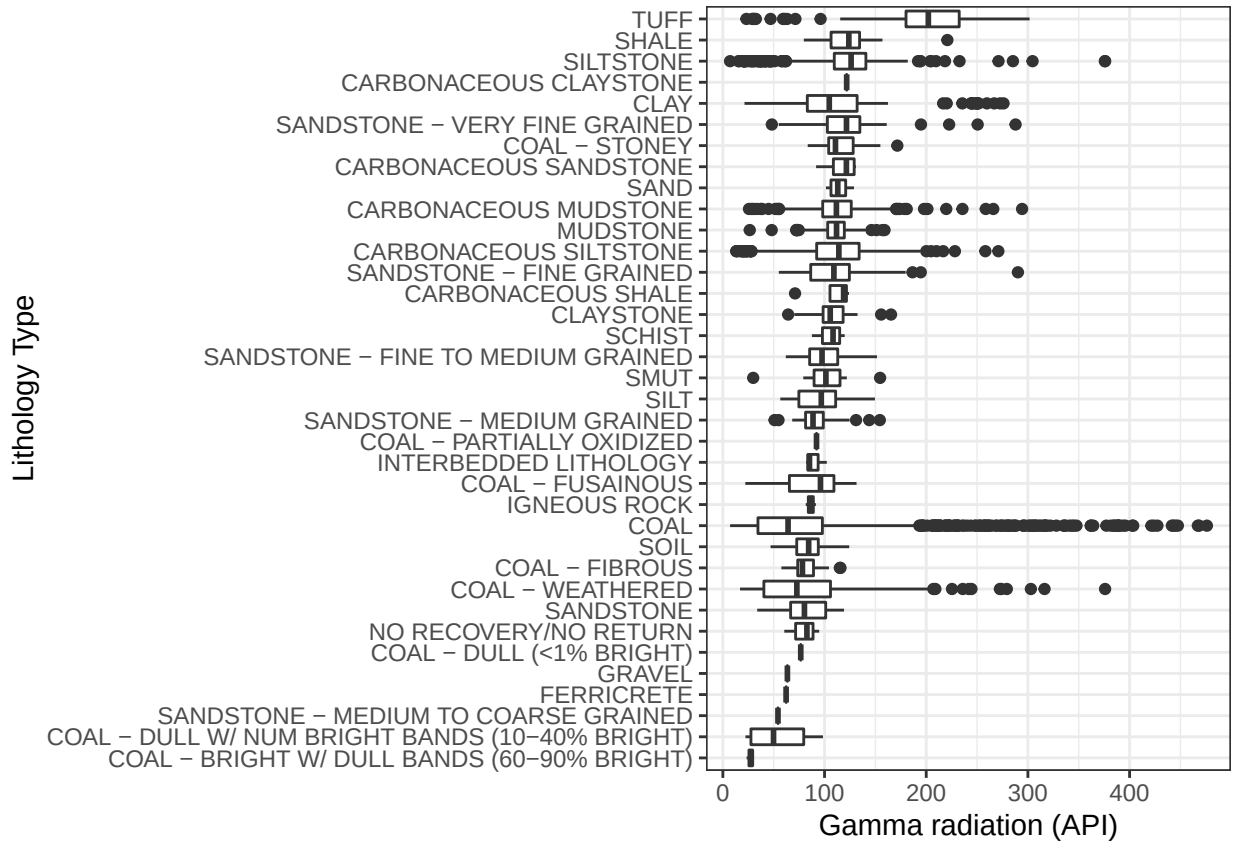

Figure 1: Boxplots of gamma log values by lithology type

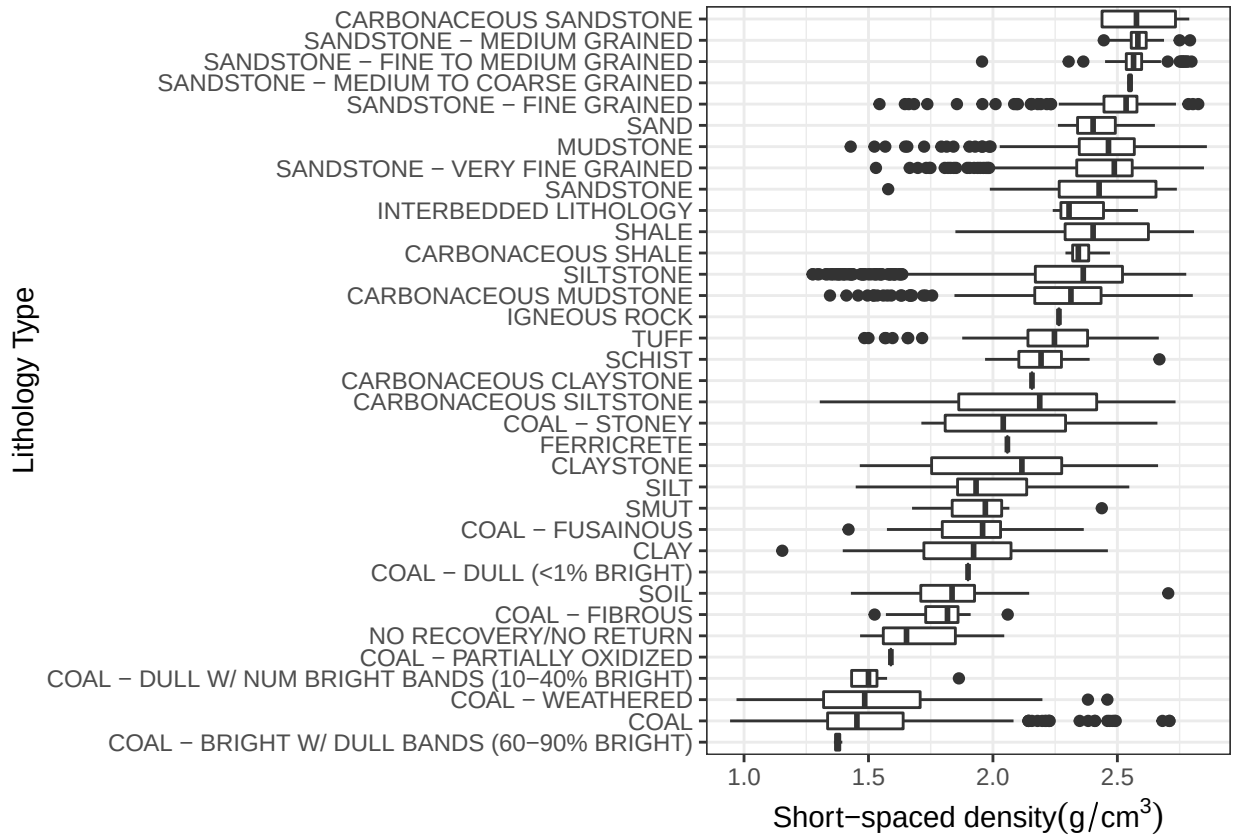

Figure 2: Boxplots of short-spaced density log values by lithology type

Both sets of boxplots show that there are a significant number of outliers for many of the rock types, but in the gamma boxplots, one rock type in particular stands out – coal. There are a lot of gamma readings for coal that have extremely high values; far higher than should be the case. This is most likely due to a phenomenon known as attenuation. Attenuation occurs when faulty readings are recorded by well logging equipment when the thickness of the piece of ground the reading is recorded for is less than 25cm. This can be confirmed by plotting thickness against gamma readings for coal, which is seen in Figure 3. Figure 3 confirms the relationship between low thickness and high gamma readings for coal. For the validity of the analysis, it is important that these attenuated observations are dropped from the dataset.

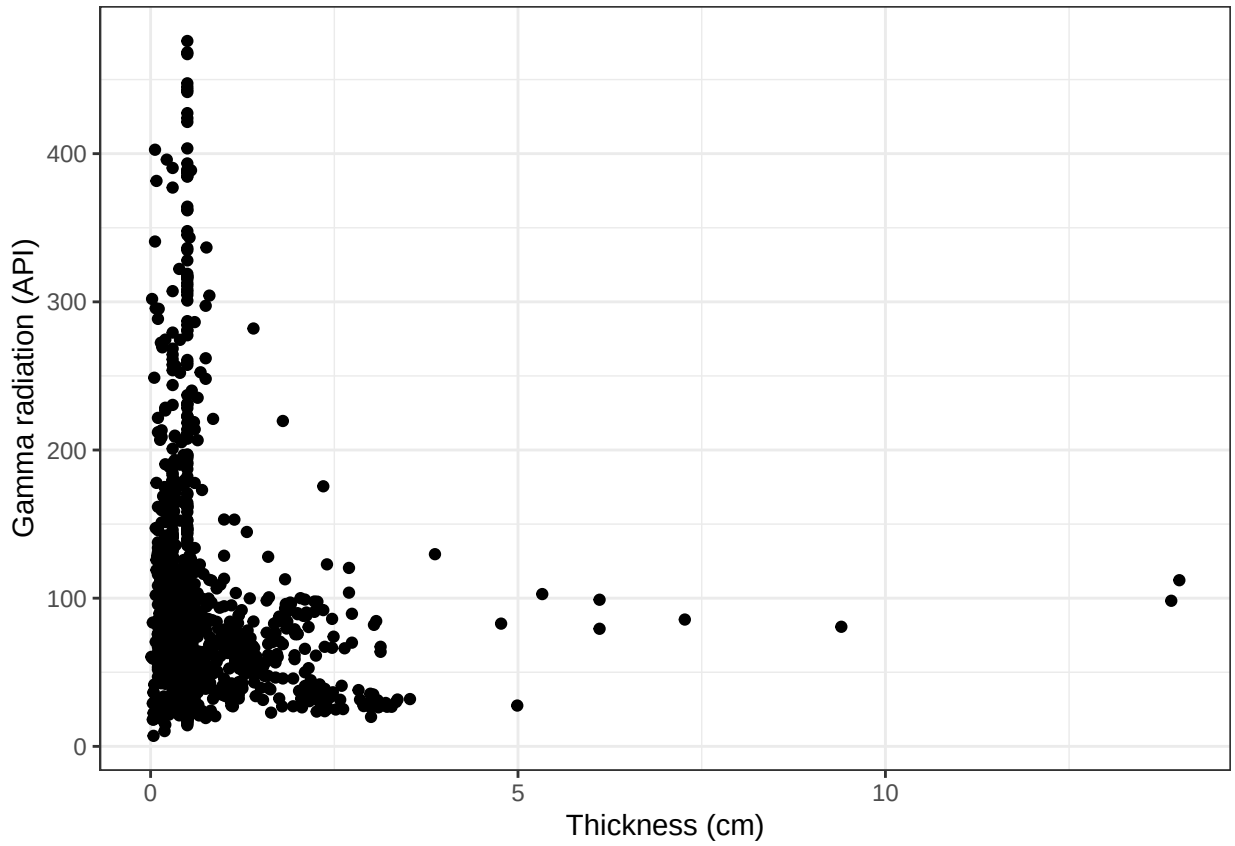

Figure 3: Plot of gamma log values against thickness, for coal

In terms of the rock types themselves, it can be seen that some groups have many observations, whereas others have very few. Leaving the groups as they are would prove problematic for predictive modelling, as some have so few observations. A lot of these rock types can be grouped together on the basis of gamma and density logs, as well as their general class (i.e. coal, siltstone, sandstone, etc...). For instance, all coal types generally have lower gamma and density readings than other rock varieties, and sandstones appear to have higher densities than other rock types. This is an important finding; gamma and density logs can be used to create larger groups which are more convenient for modelling purposes, whilst respecting that different gradations of the same rock (e.g. fine sandstone, carbonaceous sandstone, etc...) belong together, so as to make the classes interpretable. These boxplots also have the added benefit of confirming the predictive power of gamma and density logs for lithological classification.

To finalise the creation of larger lithological classes for the modelling stage, it is important to examine a cross-plot of median gamma readings against median density readings for each rock type. This cross-plot can be seen in Figure 4.

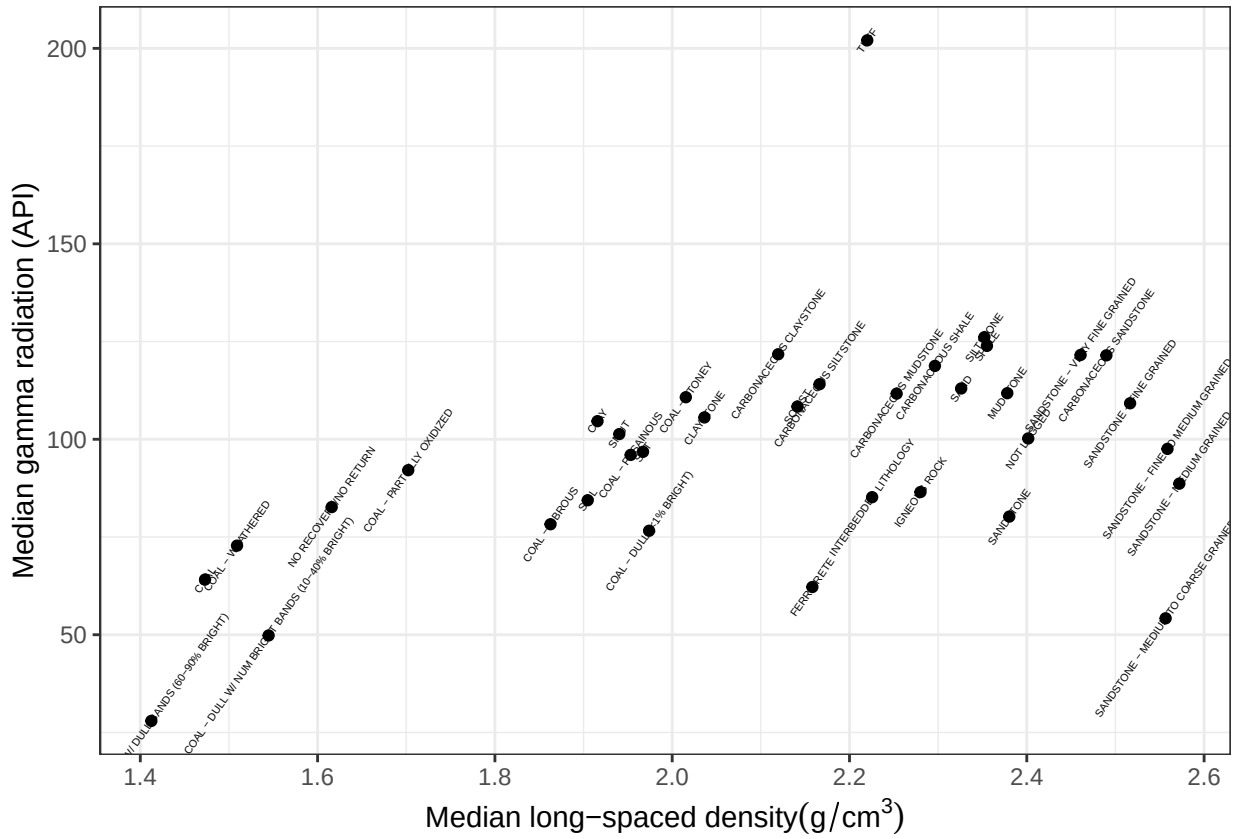

Figure 4: Cross-plot of all lithology types: median gamma vs. median long-spaced density

In this cross-plot, it can be seen that coal types cluster together in the bottom left corner of the plot, inferior coals (e.g. fibrous coal, fusainous coal, smut, etc. . .) cluster together towards the centre, and sandstones cluster together on the far-right of the plot. Rock types such as claystones, mudstones and siltstones are mixed together, which is somewhat unavoidable. This plot provides a strong basis on which to group the classes together; it was determined that these groups could be combined together into eight larger, interpretable classes: coal, inferior coal, soil, claystone, mudstone, siltstone, sandstone and tuff. These classes are used in the modelling stage of the report; their separability can be confirmed by examining a cross-plot of the eight classes, which is seen in Figure 5, which also suggests that the predictive power of the gamma and density logs is retained. It should also be noted that rock types with very small numbers of observations, observations that did not have rock types recorded, and rock types not easily categorised were dropped from the dataset.

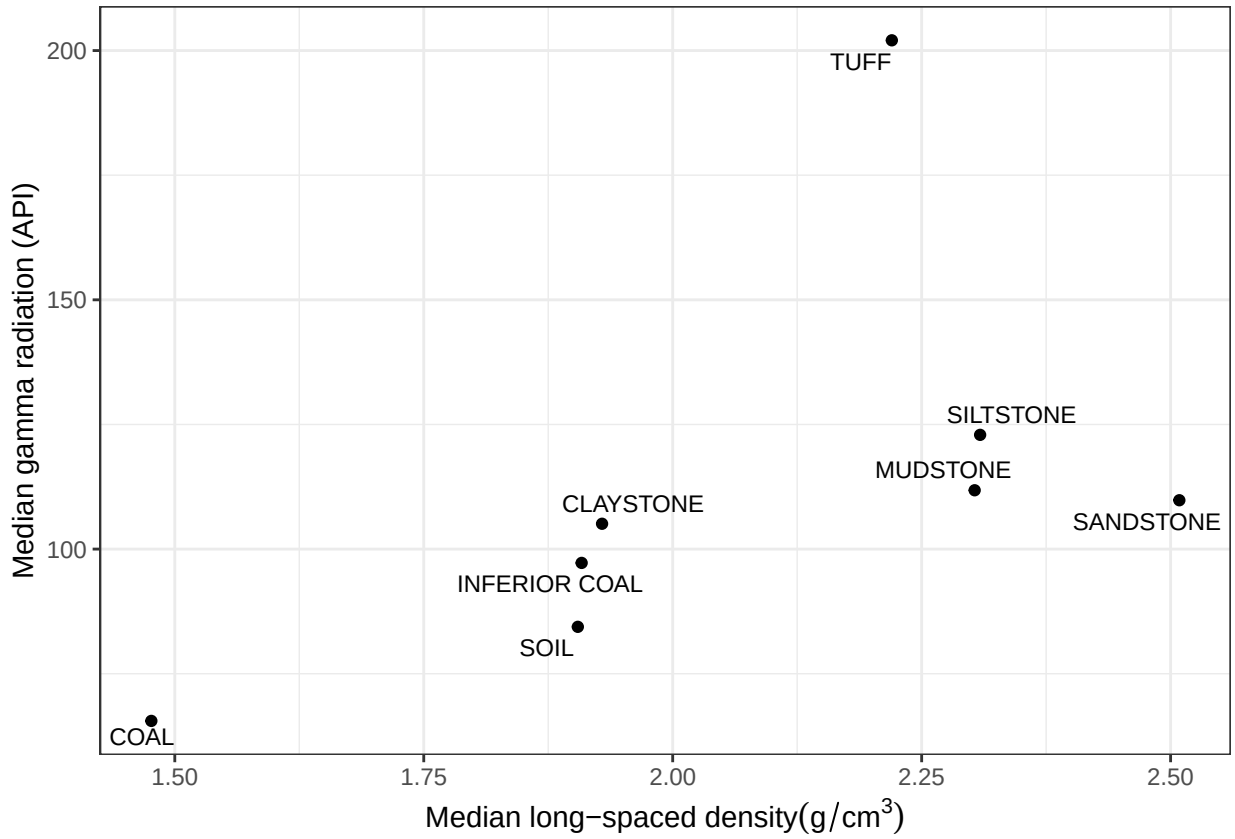

Figure 5: Cross-plot of lithology super groups: median gamma vs. median long-spaced density

The final step in the EDA for the lithological classification dataset was to confirm whether the incomplete geophysical log variables have any predictive power on the eight classes to be used in the modelling stage. This was again done with the use of box-plots, which can be seen in Figures 6 and 7. Caliper density was not assessed as it captures the same information as long- and short-spaced density logs.

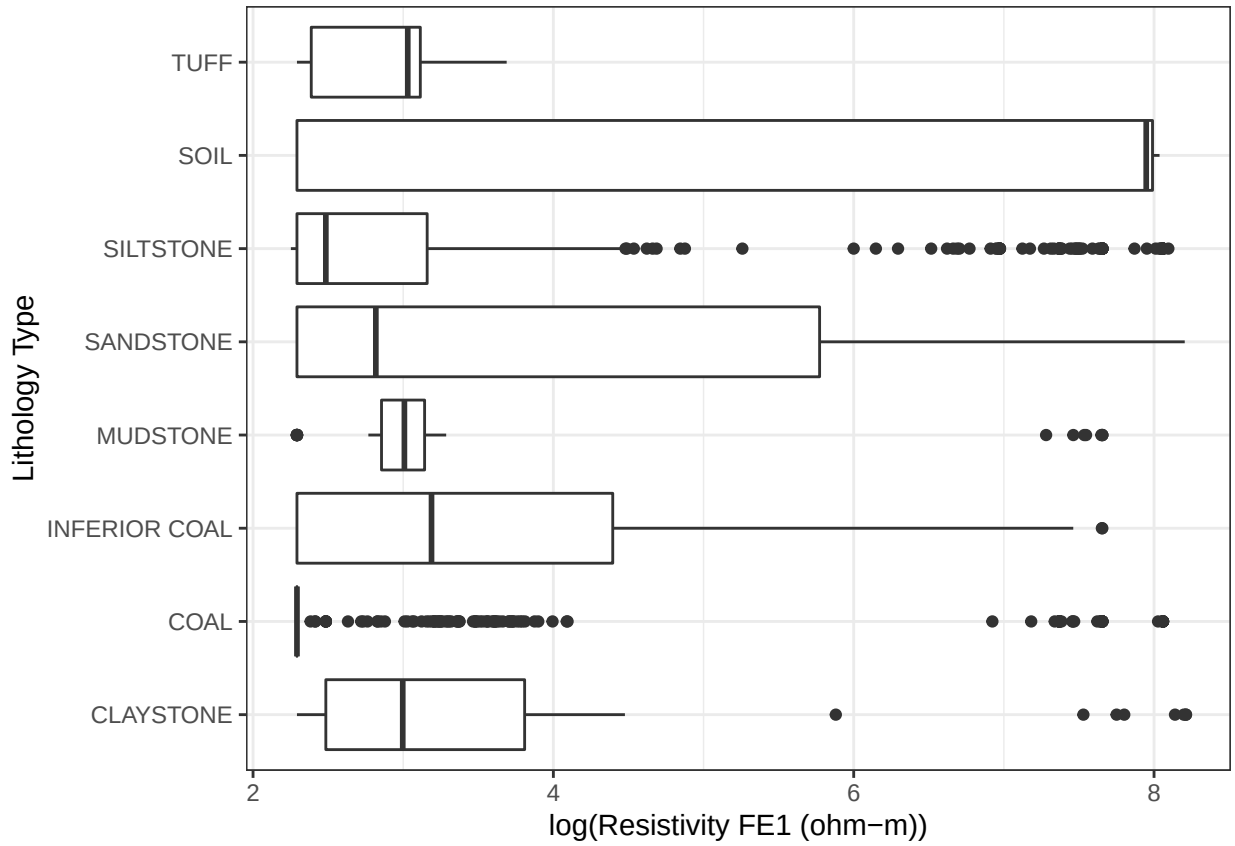

Figure 6: Boxplots of resistivity log values for every lithology super group

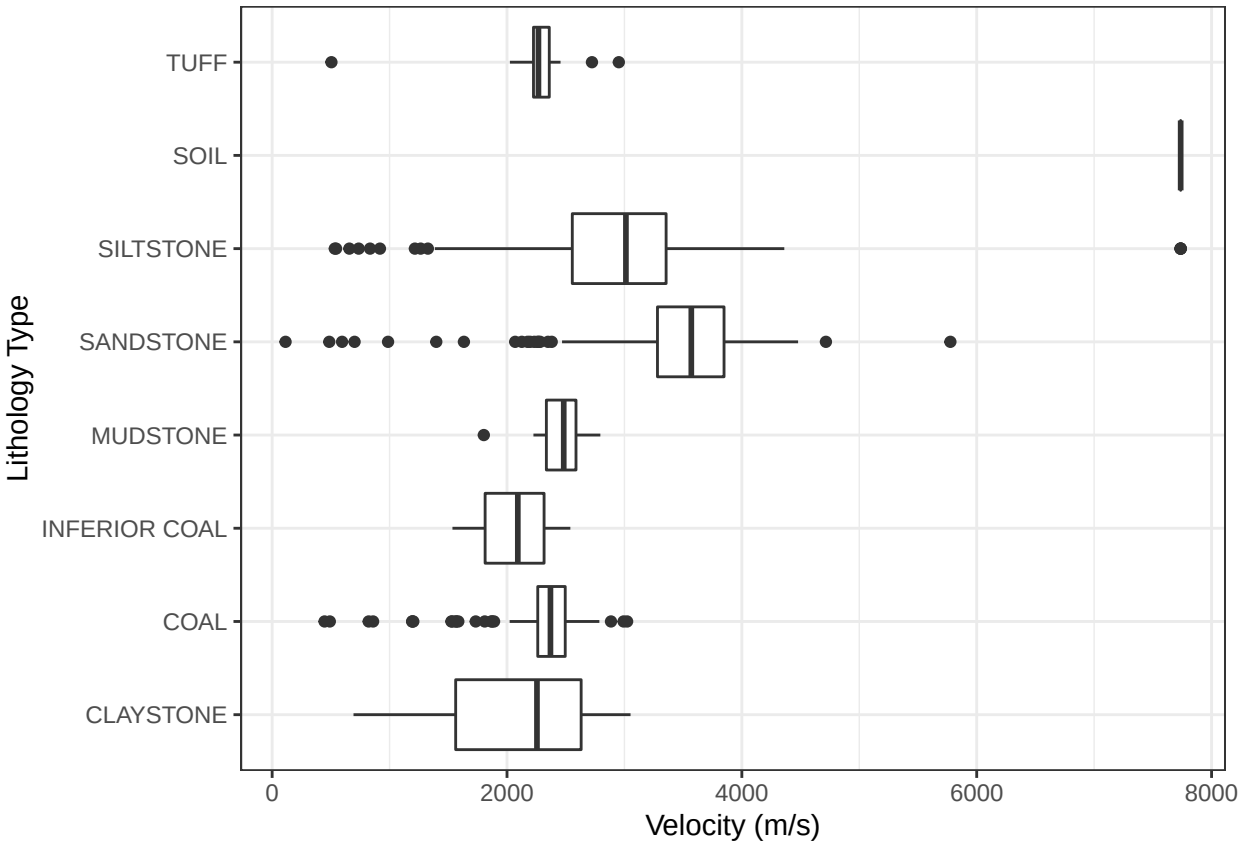

Figure 7: Boxplots of velocity log values for every lithology super group

Both boxplots show that neither resistivity nor velocity has any significant predictive power for lithological classification, given the data present. This, combined with the incompleteness of the data for the variables, means that these variables, along with caliper density, were not used in the modelling stage of the report. Based on the findings of the EDA, only gamma, long- and short-spaced density logs were retained for predictive modelling. Additional to this, any missing values for those variables were removed. All observations with thickness of less than 25cm were also removed to deal with the attenuation issue. Initial rock types were grouped together into eight larger classes, with rock types with very small numbers of observations, or that were difficult to categorise, being dropped from the dataset.

Before modelling could begin, the remaining outliers were removed from the dataset. The EDA showed evidence of outliers for both gamma and density logs in most classes. Furthermore, visual inspection of the classes showed that there were observations for most classes that had gamma and/or density readings that were not correct, given the different classes should exhibit gamma and density log properties within a certain range of values. Therefore, Mahalanobis distance was used to remove outliers. Mahalanobis distance is a measure of the distance between a point and its distribution (Mahalanobis, 1936), and is suited to multivariate outlier removal as it can be calculated in a p-dimensional space. It is also a more elegant solution to outlier removal than alternative methods (e.g. using the interquartile range). Outliers are removed on the basis of p-values from a hypothesis test that uses the distances to determine whether or not a point is part of a particular p-dimensional cluster. This was done for each of the eight classes, using a rejection rule of $p<0.05$. This is on the liberal side for outlier detection; however, due to the desire to uniformly restrict the clusters to plausible ranges of geophysical log values as much as possible, it was decided to use this rule. Figure 8 below shows that the same degree of separation between the lithology super classes was maintained after outlier removal and the preceding alterations to the dataset. The final dataset consists of 3,522 observations, with 3 predictor variables and 1 response variable used for the modelling.

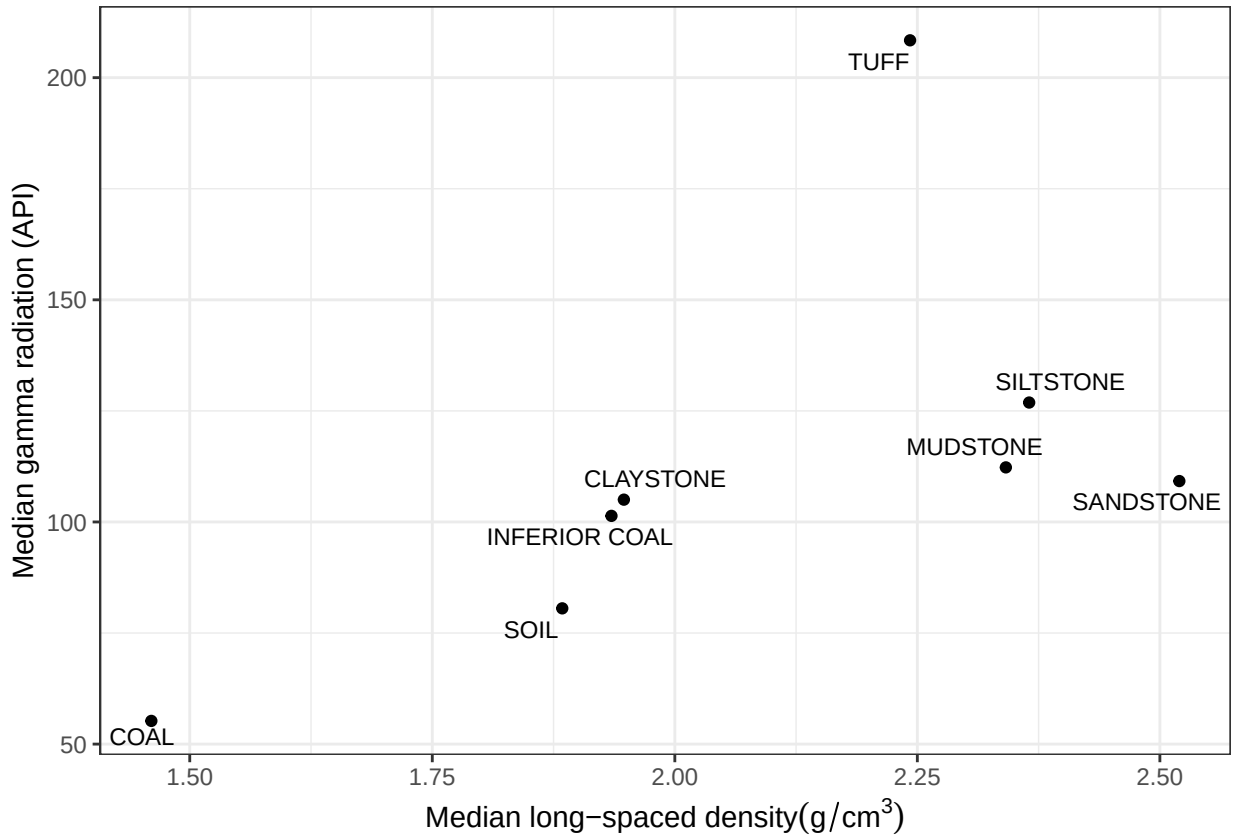

Figure 8: Cross-plot of lithology super groups post-outlier removal

Coal quality prediction

As the coal quality prediction dataset was much smaller and much more complete, less work was needed to prepare it for modelling. Again, descriptive statistics were calculated on the dataset, for each response variable for each seam; these statistics can be seen in the appendix. The descriptive statistics did not strongly indicate the presence of outliers. Scatterplots were also produced to check whether the geophysical log variables had a relationship with the response variables. These are shown in Figures 9 and 10, and confirm that the density log data exhibits strong correlation with ash, fixed carbon and volatile matter content, whilst the gamma log data exhibits somewhat weaker but still significant correlation. The figures also confirm that the density and gamma logs are not strongly related to moisture content.

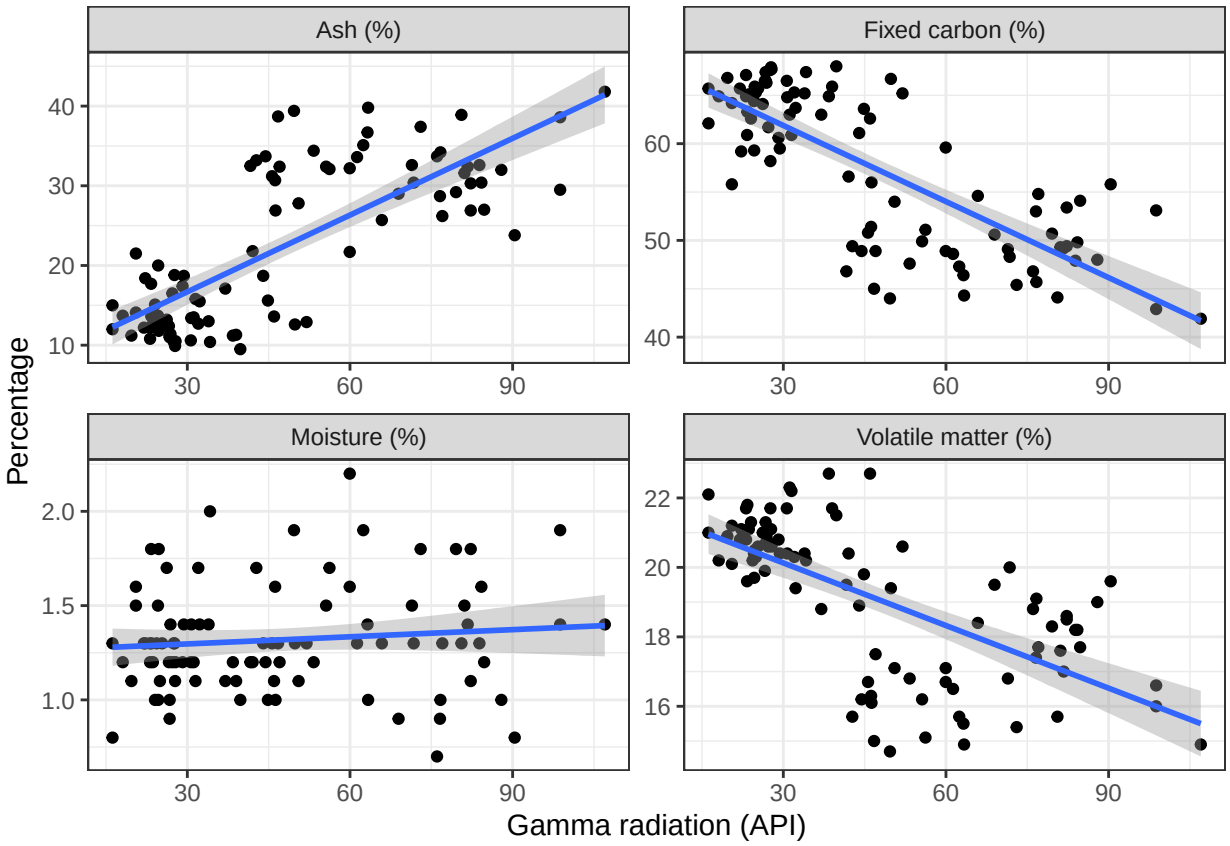

Figure 9: Gamma log plots faceted by coal quality measure

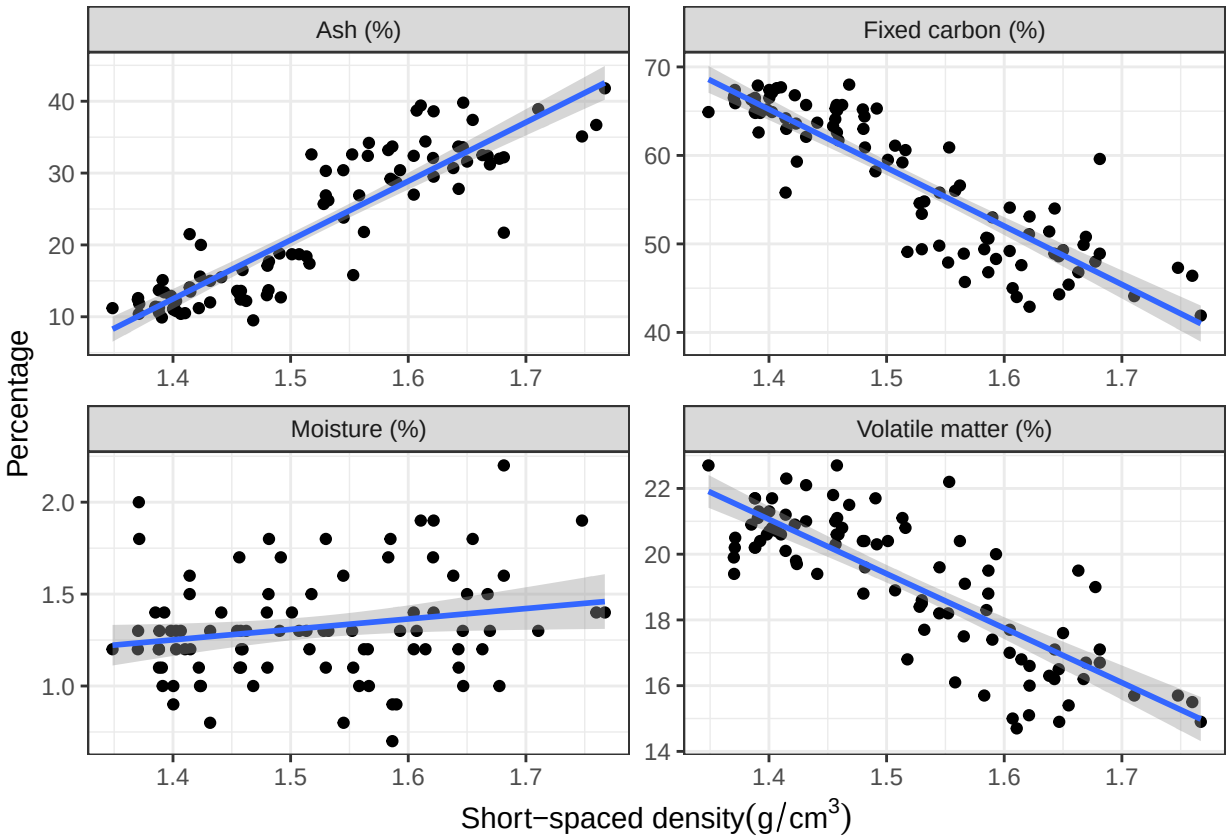

Figure 10: Short-spaced density log plots faceted by coal quality measure

In the interests of consistency and due to lack of completeness of other predictor variables in the dataset, it was again decided to only use gamma and density log readings as predictor variables in the modelling. This is fine because, as was already explained, gamma and density log variables have the most predictive power with regards to the coal quality parameters examined in this analysis. Lastly, it should be noted that for a few of the boreholes and seams, there is more than one set of coal quality measure values. This is due to there being more than one sample from a given seam from a given borehole in these cases. As these values do not appear to significantly affect the relationship between the geophysical log variables and the response variables (as evidenced by the visualisations in Figures 9 and 10), it was decided to leave these observations in the final dataset. The final dataset consists of 92 observations, with 3 predictor variables and 4 response variables used.

Modelling Methodology

The explanations of the modelling methodology featured here are predominantly sourced from Introduction to Statistical Learning, 2nd edition (James et al., 2021).

Lithological classification

For the lithological classification problem, the dataset was split into training and test sets (80%-20% split). Due to the imbalance in class sizes, it was necessary to balance the training set before analysis could begin using the supervised learning methods. This was done using the SCUT algorithm from the scutr package (Ganz, 2021); classes were randomly over/undersampled, resulting in equally-sized classes. The cluster analysis was also conducted on the test set.

10-fold cross validation (CV) was used for tuning the random forest (RF), gradient-boosted tree (GBT) and support vector machine (SVM) methods. The RF, GBT and SVM modelling was all done using the tidymodels package (Kuhn & Wickham, 2020). In each instance, the model with the highest F-measure score was selected as the best model, and was subsequently used to predict the test set observations. Each model was compared to each other on the basis of accuracy, specificity, precision, recall and F-measure, as well as their respective confusion tables. For the cluster analysis, interpretations of the clusters were made, and accuracy metrics were calculated on the basis of these interpretations.

1. Random forests

The first method used for the lithological classification problem was random forest (RF). RF is a special form of bagging. Bagging is an ensemble method that involves bootstrapping a training set, building decision trees on each bootstrapped sample, and obtaining predictions for the response from each tree. In a classification problem, the overall prediction is a majority rule of the predictions from each tree; in a regression problem, it is the average of the individual predictions. This is designed to solve the weak learning problem associated with single decision trees by reducing the variance of the response. The issue with bagging, however, is that in considering all predictor variables at each split, the individual trees can become highly correlated. This defeats the purpose of bagging, as an aggregate prediction derived from a series of correlated trees will be highly variable.

RF solve this problem by only considering a random subset of predictors at each split. This decorrelates the individual decision trees, leading to more robust aggregate predictions by making the average/majority of the resulting trees less variable. RF are convenient for a number of reasons. First, if one is concerned with merely prediction, correlation between predictors does not affect the result; this is convenient for this analysis as the long- and short-spaced density logs are highly correlated. Second, RF do not overfit the training data. Third, RF can be extended to multi-class classification and multivariate regression problems. Fourth, CV is not strictly necessary for RF models as parameters can be tuned using the out-of-bag error from the bootstrapped samples. For this problem, however, 10-fold CV was used to tune the number of predictors randomly sampled at each split, the number of trees contained in the ensemble, and the minimum number of data points in a node required for the node to be split further.

2. Gradient boosted trees

Gradient boosted trees (GBT) is an alternative ensemble method to RF. A GBT classification model works slightly differently to a GBT regression model, but only initially. First, the log odds of the response is taken, and converted into a probability. This probability is treated as the initial prediction for every instance in the training set. For all of these instances, the residuals are calculated (observed probability – predicted probability), and a decision tree is fitted to them to predict the residuals. These predictions are used to update the response predictions. The rate at which this updating occurs is controlled using a learning parameter. This process is repeated hundreds of times, with trees that are sequentially grown using information from the previous tree. The idea behind GBT is that you slowly improve the response predictions in areas where they do not perform well.

Like RF, GBT are extendable to multi-class classification problems, and correlation between predictors does not affect the result. However, GBT have some drawbacks compared to RF; GBT can technically overfit the training data, but it is unlikely, and CV is necessary for tuning the parameters. Here, 10-fold CV was used to tune the number of trees, the tree depth, and the learning parameter.

3. Support vector machines

Support vector machines (SVM) are a different type of supervised learning method to ensemble methods such as RF and GBT. They are an extension of support vector classifiers (SVC), which themselves are a less restrictive version of maximal margin classifiers (MMC). MMC perfectly classify data through the use of a hyperplane which is equidistant between margins that are defined by support vectors (i.e. data points of each class closest to the hyperplane). Often, classes cannot be perfectly separated; SVC are useful here as they

allow some points to lie within the margins, or even be misclassified. How much misclassification they allow is determined by a cost parameter.

However, SVC can only be used for linear classification problems. For non-linear classification problems, SVM are helpful. SVM work by expanding the feature space, turning a two-dimensional non-linear classification problem into a p-dimensional linear one. That is, classes that may not be able to be linearly separated in two dimensions may be able to be in p dimensions. SVM do this through the use of kernels. There are many different types of kernels that can be used; a common one is the radial kernel, which is used in this analysis. Here, 10-fold CV was used to tune the cost parameter, as well as the positive number for the radial basis function. It was also necessary in this case to standardise the predictor variables before proceeding with the analysis.

4. k-means clustering

The last method used for lithological classification was k-means clustering. Unlike the other three methods, k-means clustering is an unsupervised classification method, and does not require a train/test split or CV. This method involves selecting a pre-determined number of centroids, and assigning observations to them in a way that minimises some measure of distance. The subsequent clusters are then interpreted as to what classes they may represent. There are several algorithms that do a reasonably good job at approximating the assignment process. In R, the algorithm used is the Hartigan-Wong algorithm.

The Hartigan-Wong algorithm (Hartigan & Wong, 1979) works as follows. A number of centroids k is chosen, all points are randomly assigned to a centroid, and the centroids are calculated as the means of their assigned points. Then, the sum of squared distances of each datapoint to each centroid is calculated, and a datapoint is assigned to the centroid with which it has the smallest sum. The centroids are then recalculated again, and the process is repeated until no more datapoints change centroids. In this analysis, as for any cluster analysis when the predictor variables are in different scales, it was necessary to standardise the data before analysis.

Coal quality prediction

For the coal quality prediction problem, the data was split into training and test sets (75%-25% split). Due to the small size of the dataset, leave-one-out CV (LOO-CV) was used to tune the model parameters (in the case of the RF models, OOB error was used). The optimal models in each case were then used to predict the test set data, with the models being compared to each other on the basis of root-mean squared error (RMSE), mean absolute error (MAE), and R-squared (RSQ).

1. Random forests

All relevant points regarding RF in the lithological classification section apply here as well. The difference is that this problem deals with multiple response variables that should add up to 100. Three different methods were used in this analysis. The first was multivariate RF (multi-RF), which predicted all of the response variables simultaneously. The second was "compositional" RF (comp-RF), where the response variables underwent isometric log-ratio transformation, with the resulting vectors being individually predicted and then back-transformed into the original response variables. This was done using the compositions package (van den Boogaart et al., 2022). The third was the prediction of each untransformed response variable separately (indi-RF). All modelling here was done using the randomForestSRC package (Ishwaran & Kogalur, 2022).

2. Multivariate principal components regression

Multivariate principal components regression (multi-PCR) works the same way as a multivariate regression, except principal components (PCs) are used as regressors, instead of the original predictor variables. This is done when the original predictors are highly correlated. Multi-PCR deals with this problem as PCs are uncorrelated with each other by design. The only risk with multi-PCR is that, in not using all of the PCs, one risks losing information that may be helpful for predictive accuracy. All modelling in this section was done using the pls package (Liland et al., 2021).

Results

Lithological classification

Training data performance metrics

For the supervised learning methods, the best performing RF model randomly sampled 1 predictor per split, had 718 trees, and required a minimum of five data points in a node for it to be split further. The best performing GBT model had 1753 trees, a tree depth of 5, and a learning parameter of 0.009 (3dp). The best performing SVM model had a cost parameter of 8.867 (3dp), and a positive number for the radial basis function of 0.120 (3dp). The tables containing the performance metrics for each model on the training data are included below.

Table 2: Training set performance metrics for supervised classification models

Metric	RF	GBT	SVM
Accuracy	0.845	0.845	0.685
F-measure	0.840	0.840	0.676
Precision	0.841	0.841	0.687
Recall	0.845	0.844	0.685
Specificity	0.978	0.978	0.955

For these optimal models, robustness can be checked by examining whether the accuracy metrics are similar for each cross-validation fold. These plots are included below. All three plots demonstrate a high-level of consistency in accuracy across the cross-validation folds. This indicates that each of the optimal models is fairly robust.

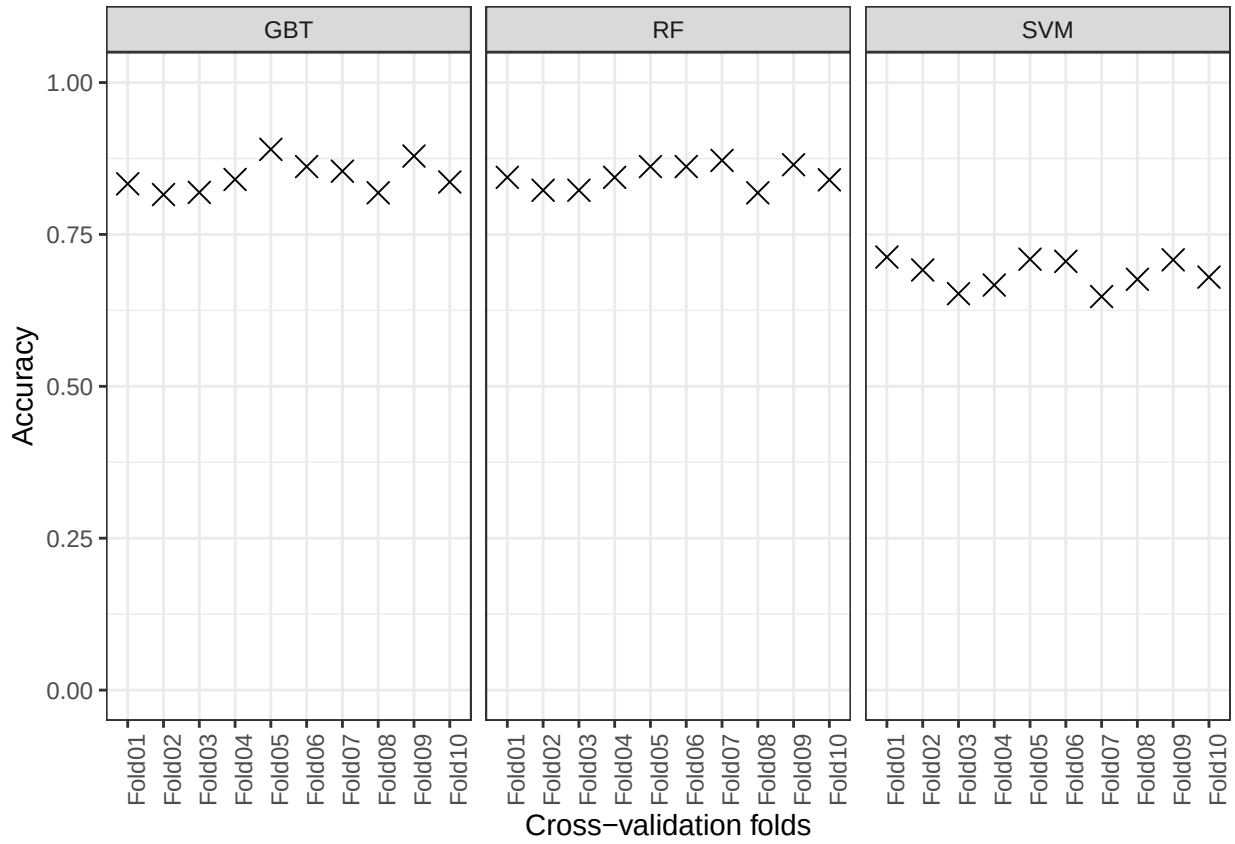

Figure 11: Plot of accuracy by cross-validation fold for each model

Test data performance metrics

For the supervised learning methods, the test data performance metrics are included in the tables below. The confusion tables for the three models on the test data are included in the appendix. It should be noted that weighted macro averaging was used to calculate the metrics, meaning that accuracy and recall are identical.

Table 3: Test set performance metrics for supervised classification models

Metric	GBT	RF	SVM
Accuracy	0.718	0.737	0.701
Specificity	0.950	0.954	0.960
Precision	0.760	0.774	0.795
Recall	0.718	0.737	0.701
F-measure	0.732	0.748	0.738

For the unsupervised learning method (k-means), it was possible to calculate performance metrics on the test data after designating each cluster as a lithotype on the basis of interpretation. The test data performance metrics are included in the table below, and the confusion table for this method is included in the appendix.

Table 4: Test set performance metrics for k-means clustering approach

Metric	kmeans
Accuracy	0.473
Specificity	0.936
Precision	0.697
Recall	0.473
F-measure	0.532

Coal quality prediction

For the RF methods used, the models were tuned using OOB training error. For multi-RF, the optimal model randomly sampled three predictors per split (mtry), had 500 trees, and required a minimum of 8 datapoints in a node for it to be split further (min). For comp-RF, the model for the first vector had 200 trees, mtry = 3 and min = 1, the model for the second vector had 400 trees, mtry = 1 and min = 7, and the model for the third vector had 100 trees, mtry = 3 and min = 8. For indi-RF, the model for ash had 100 trees, mtry = 3 and min = 1, the model for volatile matter had 300 trees, mtry = 3 and min = 1, the model for moisture had 500 trees, mtry = 3 and min = 20, and the model for fixed carbon had 100 trees, mtry = 2 and min = 5.

For multi-PCR, LOO-CV was used to tune for the number of PCs that should be included in the model. It was decided to use just the first PC, as this model had generally slightly lower error on the training data than the model with the first and second PCs.

Test data performance metrics

For the RF models and multi-PCR, the test data performance metrics are included in the tables below. Observed vs. predicted plots for each coal quality parameter are included in the Appendix.

Table 5: Test set root-mean squared error results for coal quality models

Model	Ash	Volatile Matter	Moisture	Fixed Carbon
multi-RF	3.41	0.79	0.29	3.32
multi-PCR	3.38	0.87	0.29	3.29
comp-RF	3.44	1.04	0.30	3.26
indi-RF	3.82	0.86	0.29	3.67

Table 6: Test set mean absolute error results for coal quality models

Model	Ash	Volatile Matter	Moisture	Fixed Carbon
multi-RF	2.78	0.63	0.24	2.71
multi-PCR	2.69	0.71	0.23	2.76
comp-RF	2.77	0.81	0.25	2.58
indi-RF	2.96	0.66	0.24	2.86

Table 7: Test set R-squared results for coal quality models

Model	Ash	Volatile Matter	Moisture	Fixed Carbon
multi-RF	0.90	0.86	0.07	0.86
multi-PCR	0.90	0.83	0.10	0.87
comp-RF	0.89	0.76	0.03	0.86
indi-RF	0.87	0.83	0.11	0.82

Discussion

Lithological classification

Supervised models

For the lithological classification problem, the RF and GBT models displayed near-identical performance on the training data, each with an average accuracy across the CV folds of 84.5%. The SVM model performed significantly worse on the training data across all metrics, with average accuracy across the CV folds of 68.5%. However, the similarity in the performance of all three models on the test data was much closer. The RF model had the highest accuracy, at 73.7%, followed by the GBT model with 71.8%, with the SVM model exhibiting accuracy of 70.1%. All three models demonstrate similar results for specificity, with the SVM exhibiting marginally higher precision. As a result, the RF model exhibits a marginally higher F-measure score, followed by the SVM and GBT models; this is due to the F-measure being composed of precision and recall.

An examination of the confusion tables for each model shows that recall by class for the RF and GBT models are very similar, with RF displaying identical or marginally better recall for each class than GBT. The SVM has notably worse recall for the coal and sandstone classes, but better recall for the siltstone class, as well as the smaller classes such as claystone, inferior coal and soil. Taking this information into account, it appears that whilst the RF model is the most accurate of the three, the SVM model exhibits a slightly more balanced performance across all classes. It should also be noted that the SVM model displays marginally better precision on the larger classes than the other two models, which accounts for its superior overall precision. All models predicted the coal class with over 80% accuracy; however, if one is most focused on accurately predicting the presence of target minerals such as coal, the RF and GBT models outperform the SVM approach in this type of geological environment. In general, all three models are good at predicting highly distinctive classes such as coal and tuff, but are poorer at predicting classes that overlap and intersect. Overall, given the complexity of the classification problem, and the data available, the performance of the three models on the test data is quite impressive. This demonstrates that all three models could be used to predict lithologies from geophysical log data in an industry setting, and that the methods used here could be extended to a wider variety of geophysical and geochemical contexts.

Unsupervised model

Compared to the supervised learning models, the performance of the k-means clustering method was poor. Given the interpretation of the clusters made, the clustering approach displayed an accuracy on the test data of just 47.3%. The cluster analysis was re-run using principal components as the predictors, to check that the correlation between the short- and long-spaced density logs was not causing poor performance. However, doing this resulted in no improvement in performance. Fundamentally, the poor performance is due to the fact that the Hartigan-Wong algorithm, in trying to minimise within-cluster sum of squared distances, is unable to accurately describe the true relationships in the data. The true clusters significantly overlap and intersect each other; the Hartigan-Wong algorithm finds somewhat reasonable approximations of these clusters in some cases, but in other cases the clusters it produces are a mix of different lithologies that do not belong together. For this reason, it is not surprising that k-means clustering performed poorly on the test data, and compared to the supervised learning models, this approach would be a poor method to use in industry for lithological classification.

Coal quality prediction

For the coal quality prediction problem, the best performing models overall were multi-RF and multi-PCR. Both models, however, had slightly varied performance on the different coal quality parameters. For instance, multi-RF performed best on volatile matter, with RMSE = 0.79, and MAE = 0.63, as opposed to multi-PCR (RMSE = 0.87, MAE = 0.71). However, multi-PCR performed better on ash (RMSE = 3.38, MAE = 2.69), as opposed to multi-RF (RMSE = 3.41, MAE = 2.78). Both models explained a large degree of the variance in the data for three of the four parameters; the exception was moisture. This is not surprising, as moisture content in coal does not exhibit any relationship to the geophysical log variables used. This pattern is also seen for the other two models.

For the other two models, overall comp-RF outperformed indi-RF, but performed worse than the other two models. This is largely due to its relatively poor performance in predicting volatile matter content; in fact, comp-RF is the best performing model in predicting fixed carbon content. Indi-RF exhibits the worst performance of the methods examined, mainly due to poor performance on the ash and fixed carbon content parameters. It should also be noted that, unlike the other three methods, the predictions from this approach do not add up to 100. Sometimes this approach can lead to lower test error, but the predictions are not truly valid as the four parameter values should always add to 100. Overall, the performance of the three multivariate/compositional models is quite impressive; the predictions are a highly accurate approximation of the true data. The results show that it is feasible to use geophysical log variables in multivariate and compositional models so as to accurately predict coal quality parameters that add up to 100, and that these methods could subsequently be used in industry. These methods could also be extended to other target minerals in the mining industry where quality assessment is concerned.

Dashboard

The details of the three supervised models for lithological classification, and the four coal quality prediction models, were shown in a dashboard designed as a prototype for industry use. The dashboard was developed predominantly through the use of the flexdashboard (Iannone et al., 2020) and shiny (Chang et al., 2021) packages in R. For lithological classification, the dashboard includes a visualisation of the predictions, a confusion table and performance metric table for the test data results, and a plot of accuracy on the training set CV folds. For coal quality prediction, the dashboard includes a visualisation of the predicted values against the actual values, and a table of performance metrics on the test data. For each section, the user is able to toggle between models using a radio button selector. The dashboard also includes a panel that explains its use.

Conclusion

The analysis found that, for lithological classification, the RF model was the most accurate at correctly identifying rock types in a complex geological environment using gamma and density log data, with an accuracy rate of 73.7%. In terms of the correct identification of coal, all three supervised models performed well, but the RF and GBT methods performed better than the SVM approach. However, the SVM model exhibited a more balanced performance across all class types, with equivalent or higher recall for all classes except coal and sandstone when compared to the RF and GBT approaches, as well as marginally higher precision on larger classes. The analysis also found that a k-means clustering approach to lithological classification using gamma and density log data performs poorly in a complex geological environment. For the coal quality prediction problem, the multi-RF and multi-PCR models were overall the best performing, with multi-RF best at predicting volatile matter content, and the multi-PCR best at predicting ash content. The comp-RF model did not perform as well overall, but was best at predicting fixed carbon content. The performance of the indi-RF model in comparison to the other three models was relatively poor, but still reasonable overall. These predictions were successfully displayed in a dashboard that serves as a prototype for industry-use.

The practical implications of this analysis are significant. Firstly, it shows that open-source gamma and density log data can be used to create relatively impressive supervised predictive models for lithological

classification in a complex geological environment. In particular, in the context of coal, all three supervised learning models have a high true positive rate; delineating coal from other rock types in an automated fashion is feasible using just gamma and density log data. This finding supports the idea that mining companies can successfully use machine learning models to classify rock types based on geophysical log data, thereby improving accuracy and saving time and money. These methods can be applied to other rock types and minerals, and can also be extended to activities such as mapping, safety analysis, and even non-mining geo-investigative activities. Second, the analysis also shows that multivariate and compositional models that use open-source gamma and density log data can predict coal quality parameters with a high degree of accuracy. This is important as it means the process of coal quality analysis can be automated, again improving accuracy, and saving time and money. These models can also be used on other minerals for quality assessment. Third, the report has demonstrated that these predictions can be displayed in a dashboard that serves as a prototype for further development for use in industry.

It is important to recognise some of the limitations of this analysis. The main limitation surrounds the data used. Using open-source data is fraught with danger regarding the completeness of datasets, especially in the mining industry, where data can often be highly fragmented. In this analysis, the incompleteness of some datasets resulted in quite a high degree of loss of observations, especially for the coal quality prediction section. This can result in analysis which is sensitive to individual or small groups of observations, and datasets which may not be truly representative of a real-world scenario. Another element of this problem can be the loss of predictors, which was also an issue with this analysis. Other geophysical log variables, had they been more complete, may have been able to improve the predictions produced. A second limitation of the analysis regards the extension of these modelling methods to other mineral contexts. Coal has highly distinctive geophysical log signatures which make it relatively easy to identify, and its quality parameters have quite strong relationships with certain geophysical log variables. The same may not be the case for other minerals. For other minerals, different geophysical log variables may have to be used, or they may not be readily distinguishable or assessable using these variables at all.

References

Al-Anazi, A., & Gates, I. (2010). On the Capability of Support Vector Machines to Classify Lithology from Well Logs. *Natural Resources Research*, *19*, 125-139.

Bhattacharya, S., Carr, T., & Pal, M. (2016). Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA. *Journal of Natural Gas Science and Engineering*, *33*, 1119-1133.

Bressan, T., Souza, M., Girelli, T., & Chemale, F. (2020). Evaluation of machine learning techniques for lithology classification using geophysical data. *Computers and Geosciences*, *139*.

Chamberlain, S., Costigan, I., Wu, W., Mayer, F., & Gelfand, S. (2021). *ckanr: Client for the Comprehensive Knowledge Archive Network ('CKAN') API*. https://docs.ropensci.org/ckanr/(website) https://github.com/ropensci/ckanr (devel).

Chang, W. et al. (2021). Shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny

Dick, M., & Maxwell, K. (2022). lastools: Tools for reading/writing and manipulating Canadian Well Logging Society (CWLS) Format V1.2 or 2.0 LAS Files as R Objects. R package version 1.0.0.

Firth D. (1999). *Log Analysis for Mining Applications*. Reeves.

Fullagar, P., Zhou, B., & Fallon, G. (1999). Automated interpretation of geophysical borehole logs for orebody delineation and grade estimation. *Mineral Resources Engineering*, *8*(3), 269-284.

Ganz, K. (2021). *scutr: Balancing Multiclass Datasets for Classification Tasks*. R package version 0.1.2, https://github.com/s-kganz/scutr.

Geological Survey of Queensland (2022). cr073851 ["Data sets"].

Ghosh, S., Chatterjee, R., & Shanker, P. (2016). Estimation of ash, moisture content and detection of coal lithofacies from well logs using regression and artificial neural network modelling. *Fuel*, *177*, 279-287.

Gordon, J., Sanei, H., & Pedersen, P. (2022). Predicting hydrogen and oxygen indices (HI, OI) from conventional well logs using a random forest machine learning algorithm. *International Journal of Coal Geology*, *249*.

Haldar, S.K. (2013). *Mineral Exploration*. Elsevier.

Hartigan, J.A., & Wong, M.A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

Iannone, R., Allaire, J., & Borges, B. (2020). Flexdashboard: R Markdown Format for Flexible Dashboards. R package version 0.5.2. https://CRAN.R-project.org/package=flexdashboard

Insua, T., Hamel, L., Moran, K., Anderson, L., & Webster, J. (2015). Advanced classification of carbonate sediments based on physical properties. *Sedimentology*, *62*, 590-606.

Ishwaran H. & Kogalur U.B. (2022). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 3.1.1.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd edition). Springer.

Jeuken, R., Xu, C., & Dowd, P. (2020). Improving Coal Quality Estimations with Geostatistics and Geophysical Logs. *Natural Resources Research*, *29*(4), 2529-2546.

Jie, H., Changchun, Z., Zhaohui, H., Liang, X., Yuqing, Y., Guohua, Z., & Wenwen, W. (2014). Log evaluation of a coalbed methane (CBM) reservoir: a case study in the southern Qinshui basin, China. *Journal of Geophysics and Engineering*, *11*(1).

Kuhn, M. et al. (2020). Tidymodels: a collection of packages for modelling and machine learning using tidyverse principles. https://www.tidymodels.org

Kumar, T., Seelam, N., & Rao, G. (2022). Lithology Prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India. *Journal of Applied Geophysics*, *199*.

Larkin, B., & Green, D. (2021). *CoalLog Manual Version 3.1 – Borehole Data Standard for the Australian Coal Industry*. Australasian Institute of Mining and Metallurgy.

Liland, K., Mevik, B., & Wehrens, R. (2022). *pls: Partial Least Squares and Principal Component Regression*. R package version 2.8-1, https://github.com/khliland/pls.

Mahalanobis, P.C. (1936). On the Generalised Distance in Statistics. *Proceedings of the National Institute of Science of India*, *2*, 45.

Maxwell, K., Rajabi, M., & Esterle, J. (2019). Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. *International Journal of Coal Geology*, *214*.

Rau, E., James, S., Breen, K., Atchley, S., Thorson, A., & Yeates, D. (2022). Applicability of decision tree-based machine learning models in the prediction of core-calibrated shale facies from wireline logs in the late Devonian Duvernay Formation, Alberta, Canada. *Interpretation (Tulsa)*, *10*, 1-45.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Siregar, I., Niu, Y., Mostaghimi, P., & Armstrong, R.T. (2017). Coal ash content estimation using fuzzy curves and ensemble neural networks for well log analysis. *International Journal of Coal Geology*, *181*, 11-22.

Sun, Y., Chen, J., Yan, P., Zhong, J., Sun, Y., & Jin, X. (2022). Lithology identification of Uranium-Bearing Sand Bodies Using Logging Data Based on a BP Neural Network. *Minerals*, *12*.

van den Boogaart, KG., Tolosana-Delgado, R., & Bren, M. (2022). *compositions: Compositional Data Analysis*. R package version 2.0-4, http://www.stat.boogaart.de/compositions/.

Wickham, H. et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. (2018). Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering*, *160*, 182-193.

Zhong, R., Johnson, R., & Chen, Z. (2020). Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). *International Journal of Coal Geology*, *220*.

Zhou, B., & Guo, H. (2020). Applications of Geophysical Logs to Coal Mining – Some Illustrative Examples. *Resources (Basel)*, *9*(2), 11.

Zhou, B., & O'Brien, G. (2016). Improving coal quality estimation through multiple geophysical log analysis. *International Journal of Coal Geology*, *167*, 75-92.

Appendix

Table 8: Table of descriptive statistics for coal quality parameters

Seam	Measure	min	max	q1	mean	q3	sd	kurtosis
LL	Ash (%)	10.4	21.5	11.100	12.395652	13.350	2.3270110	7.3253417
LL	Fixed carbon (%)	55.8	67.7	64.900	65.500000	66.650	2.3731452	9.2548887
LL	Moisture (%)	0.9	2.0	1.200	1.308696	1.400	0.2484736	0.8802729
LL	Volatile matter (%)	19.9	22.7	20.350	20.795652	21.050	0.6622814	0.7576473
LU	Ash (%)	9.5	20.0	12.600	14.795238	17.100	3.0020786	-1.1725866
LU	Fixed carbon (%)	58.2	68.0	61.100	63.347619	65.300	2.7875830	-1.0925268
LU	Moisture (%)	0.8	1.8	1.100	1.261905	1.400	0.2519448	-0.3878346
LU	Volatile matter (%)	18.8	22.7	19.700	20.595238	21.100	1.0892549	-1.0214415
VL	Ash (%)	23.8	41.8	28.700	30.790476	32.600	4.1905733	0.4938621
VL	Fixed carbon (%)	41.9	55.8	48.000	49.923809	53.100	3.7678780	-0.6589769
VL	Moisture (%)	0.7	1.9	1.000	1.290476	1.500	0.3330237	-0.9564403
VL	Volatile matter (%)	14.9	20.0	17.400	17.995238	18.800	1.2511899	-0.1691707
VU1	Ash (%)	13.5	32.5	14.350	18.900000	20.100	6.6778240	-0.4041644
VU1	Fixed carbon (%)	46.8	63.3	57.900	58.871429	62.800	5.8385827	-0.2784677
VU1	Moisture (%)	1.0	1.3	1.200	1.200000	1.250	0.1000000	-0.4285714
VU1	Volatile matter (%)	19.5	22.3	20.600	21.028571	21.550	0.9196273	-1.2840784
VU2	Ash (%)	21.7	39.8	31.875	33.415000	36.875	4.5538041	0.1072970
VU2	Fixed carbon (%)	44.0	59.6	46.150	49.080000	50.875	4.0553538	0.2373373
VU2	Moisture (%)	1.0	2.2	1.200	1.460000	1.700	0.3299123	-0.8125106
VU2	Volatile matter (%)	14.7	17.5	15.475	16.045000	16.700	0.8035939	-1.2031704

Table 9: Confusion table for random forest classification model

	CLAYSTONE	COAL	INFERIOR COAL	MUDSTONE	SANDSTONE	SILTSTONE	SOIL	TUFF
CLAYSTONE	3	10	4	6	0	6	1	0
COAL	2	249	1	0	0	1	1	0
INFERIOR COAL	0	2	2	1	0	3	0	0
MUDSTONE	6	0	0	71	10	28	0	1

	CLAYSTONE	COAL	INFERIOR COAL	MUDSTONE	SANDSTONE	SILTSTONE	SOIL	TUFF
SANDSTONE	2	0	0	8	58	43	0	0
SILTSTONE	1	1	1	6	22	122	2	0
SOIL	1	5	1	0	1	3	1	0
TUFF	1	0	0	2	0	3	0	14

Table 10: Confusion table for gradient-boosted tree classification model

	CLAYSTONE	COAL	INFERIOR COAL	MUDSTONE	SANDSTONE	SILTSTONE	SOIL	TUFF
CLAYSTONE	3	13	3	9	0	7	1	0
COAL	2	246	1	0	0	1	1	0
INFERIOR COAL	0	2	2	1	0	2	1	0
MUDSTONE	6	0	1	66	10	26	0	1
SANDSTONE	0	0	0	7	56	48	0	0
SILTSTONE	3	0	1	9	25	118	0	0
SOIL	1	6	1	0	0	3	2	0
TUFF	1	0	0	2	0	4	0	14

Table 11: Confusion table for support vector machine classification model

	CLAYSTONE	COAL	INFERIOR COAL	MUDSTONE	SANDSTONE	SILTSTONE	SOIL	TUFF
CLAYSTONE	4	17	2	0	1	13	1	0
COAL	1	215	1	0	0	0	0	0
INFERIOR COAL	5	12	3	9	3	8	0	0
MUDSTONE	3	0	1	68	11	17	0	0
SANDSTONE	0	0	0	10	50	30	0	0
SILTSTONE	1	0	1	5	26	137	0	1
SOIL	1	23	1	0	0	3	4	0
TUFF	1	0	0	2	0	1	0	14

Table 12: Confusion table for k-means clustering method

	CLAYSTONE	COAL	INFERIOR COAL	MUDSTONE	SANDSTONE	SILTSTONE	SOIL	TUFF
CLAYSTONE	1	17	0	0	0	0	0	0
COAL	0	125	0	0	0	0	0	0
INFERIOR COAL	3	56	2	0	0	2	0	0
MUDSTONE	9	4	5	29	8	65	0	1
SANDSTONE	0	0	0	50	62	41	0	0
SILTSTONE	1	0	0	15	21	97	0	1
SOIL	1	65	2	0	0	4	5	0

	CLAYSTONE	COAL	INFERIOR COAL	MUDSTONE	SANDSTONE	SILTSTONE	SOIL	TUFF
TUFF	1	0	0	0	0	0	0	13

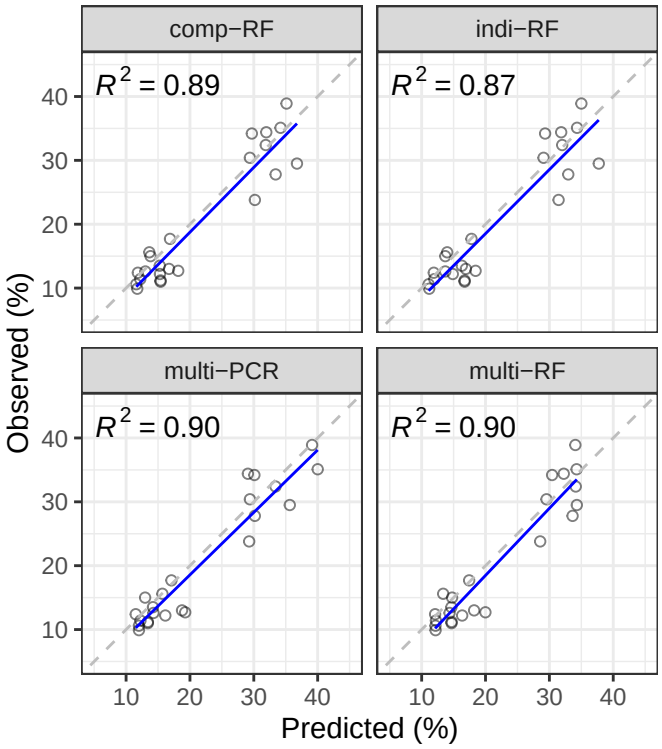

Figure 12: Observed vs. predicted plots for Ash content

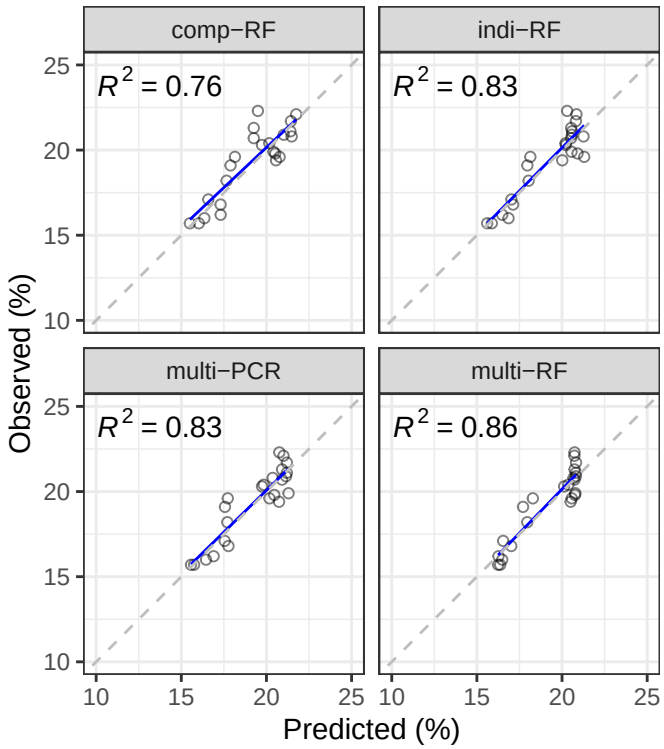

Figure 13: Observed vs. predicted plots for Volatile matter content

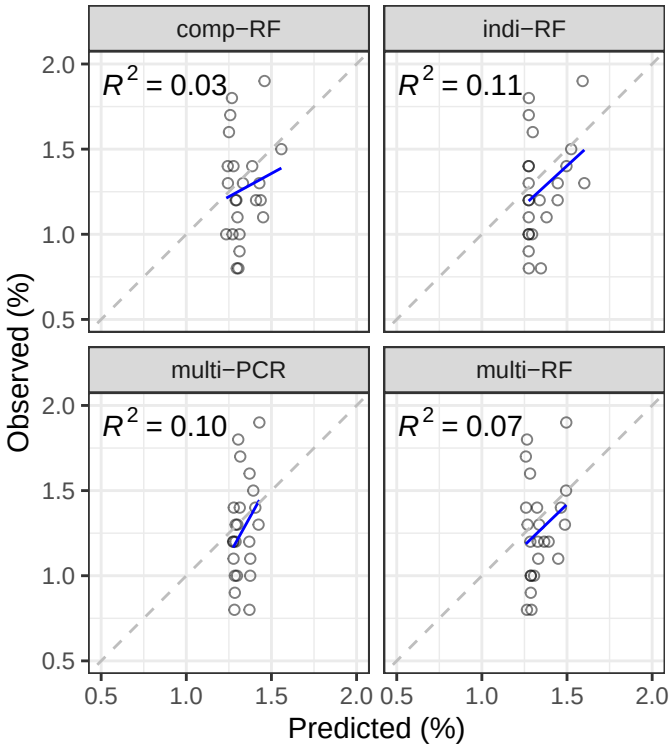

Figure 14: Observed vs. predicted plots for Moisture content

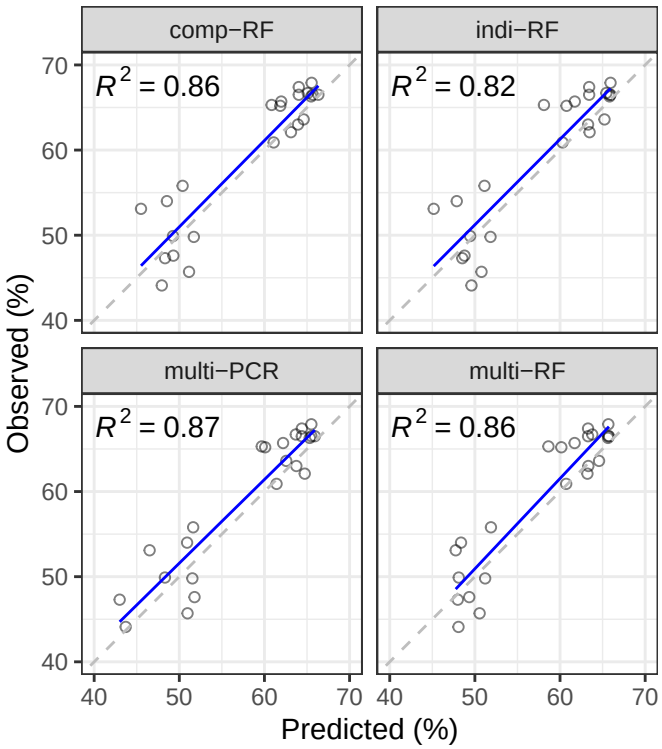

Figure 15: Observed vs. predicted plots for Fixed carbon content