# The application of machine learning methods to aggregate geochemistry predicts quarry source location: an example from Ireland

[a]Tadhg Dornan, [b]Gary O'Sullivan, [c]Neal O'Riain, [d]Eva Stueeken, [a]Robbie Goodhue

[a]Trinity College Dublin (TCD), Department of Geology, Museum building, College Green, Dublin 2, Ireland/Irish Centre for Research in Applied Geosciences (iCRAG), O'Brien Centre for Science (East), University College Dublin, Belfield, Dublin 4, Ireland
[b]UCD School of Earth Sciences, University College Dublin, Dublin 2, Ireland
[c]Unafilliated
[d]School of Earth and Environmental Sciences, University of St Andrews, St Andrews, Scotland, UK

## Abstract

Attempts using geochemical data to classify quarry sources which provided reactive rock aggregate, composed of Carboniferous aged pyritic mudrocks and limestones, which has caused structural damamge to over 12, 500 homes across Ireland have not yet succeeded. In this paper, a possible solution to this problem is found by performing machine learning models, such as Logistic regression and Random Forest, upon a geochemical dataset obtained through the scanning electron microscope energy-dispersive X-ray spectroscopy (SEM-EDS) and Laser ablation-quadrupole-inductively couple plasma mass spectrometry (LA-Q-ICPMS) of pyrite and Isotope ratio mass spectrometry (IRMS) of bulk rock aggregate, to predict quarry source location. When comparing the classification scores, the LA-Q-ICPMS dataset achieved the highest average classification score of 55.38 % for Random Forest and 67.73 % for Logistic regression based on 10-fold cross validation testing. As a result, this dataset was then used to classify a set of known unknown samples and achieved average classification accuracies of 40.30 % for random forest and 66.80 % for logistic regression, based on a systematic train-test procedure.
There is scope to enhance these classification scores to an accuracy of 100 % by combining the geochemical datasets together. However, due to the difficulty in linking pyrites analysed by SEM-EDS to those analysed by LA-Q-ICPMS, and relating a bulk rock analytical technique (IRMS) to mineral geochemistry (SEM-EDS, LA-Q-ICPMS), median values have to be used when combining IRMS (Fe, S) and SEM-EDS (TS and $\delta^{34}S$) datasets with LA-Q-ICPMS data. Therefore, if these combined datasets were used as part of an applied quarry classification system, statistically meaningful mean values taken from a near normally distributed dataset would have to be used in order to accurately represent the quarry composition.

# 1. Introduction

Between 1995 and 2007 Ireland experienced a housing boom as the number of dwellings throughout the country increased by over 88 %. (Tuohy, 2012). This unprecedented increase in construction activity led to an associated demand for construction materials, such as rock aggregate. However, much of the material delivered to these newly built housing developments was of unknown origin and not compliant with European and Irish standards, partly due to a failure to provide and/or maintain the necessary documentation (Matheson and Quigley, 2016). In many cases, a range of quarry sources may have been used as part of a single housing development and the associated records documenting the quarry source for fill in a particular dwelling was commonly absent or misidentified (Tuohy, 2012). As a result, much of the interest surrounding this aggregate material focuses on identifying the likely quarry of origin.

Due to the shared compositional and textural characteristics between many of the quarry sources, distinguishing samples by hand specimen or thin section mineralogy is both fraught with subjectivity, and agreements on quarry identification by this method are often disputed. Therefore, a quantitative method for quarry source identification is needed. This paper aims to provide such a method by using a combination of pyrite and bulk rock geochemistry and machine learning classifiers, such as logistic regression and random forest.
Pyrite ($FeS_2$) was chosen as the mineral of interest for this study as it is commonly found as a minor constituent in all six quarry sources and, although its chemistry is dominated by Fe and S, it is known to contain a wide variety of trace elements such as Ag, As, Au, Bi, Cd, Co, Cu, Hg, Mo, Ni, Pb, Pd, Ru, Sb, Se, Sn, Te, and Zn (Lehner and Savage, 2008). The mechanism by which trace elements are incorporated into the pyrite crystal structure starts with the adsorption of trace elements onto the pyrite surface, or surfaces of pyrite precursor minerals, from surrounding water column. Once adsorbed, the trace elements are incorporated into the pyrite crystal structure through a series of reaction pathways, however, the exact method by which these occurs is still up for debate (Gregory *et al.*, 2015). The degree by which trace elements are incorporated into the pyrite crystal structure is determined by a number of factors including; rate of pyrite crystallisation, nucleation and the presence of trace elements in the water column and pore waters (Gallagher, 2016)
Consequently, trace element concentrations in pyrite can vary from parts per million (ppm), to several weight per cent (wt %) for elements such as As, Co and Ni (Lehner and Savage, 2008). As a result, pyrite can exhibit strong geochemical variation through time (Gregory *et al.*, 2015), spatially, at the section scale (Gregory *et al.*, 2017) and between stratigraphic units (Sack, Large and Gregory, 2018). Therefore, pyrite provides the ideal candidate for use as part of this classification mechanism as all of the investigated quarry sources vary geographically, across the eastern Ireland, and geologically, across four different rock units.

# 2. Materials and methods

## 2.1 Samples

### 2.1.1 Sample information

A map, stratigraphic section and detailed geological descriptions of the quarry samples used in this investigation can be found in Figure 1 and Table 1. These samples are sourced from Carboniferous aged rock formations located in the eastern part of Ireland and range in composition from organic rich carbonaceous mudrocks to clean limestones. Further information regarding pyrite morphology within the sample material can be found in Dornan *et al*, 2019. Additionally, due to ongoing litigation surrounding pyritic rock aggregate in Ireland, the exact names and locations of the quarry material investigated in this study have been redacted.

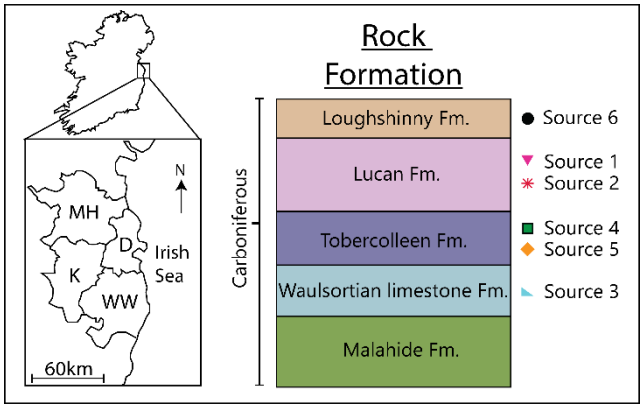



*Figure 1 (Single column, colour) Map and simplified stratigraphic section of the sampling area.*

| Formation | Lithology description | Depositional mode |
|---|---|---|
| Loughshinny Fm | Laminated to thinly-bedded, argillaceous, pyritic, locally cherty limestones interbedded with dark-grey to black shale. The limestones include argillaceous micrites and graded calcarenites (Geological Survey Ireland, 2018). | Tectonically driven collapse and drowning of the Balbriggan shelf caused a cessation in platform carbonate sedimentation and a deposition of coarse proximal basinal facies (Sevastopulo and Wyse Jackson, 2001) |
| Lucan Fm | Dark-grey to black, fine-grained, occasionally cherty, micritic limestones (Geological Survey Ireland, 2018). | Coarse-grained graded limestones with concentrations of shelly fauna at their bases are characteristic of proximal upper slope environments. The upper units are typical of more distal lower slope environments (Strogen, Jones and Somerville, 1990) |
| Tobercolleen Fm | Dark-grey, calcareous, commonly bioturbated mudstones and subordinate thin micritic limestones (Geological Survey Ireland, 2018) | The presence of well-laminated packstones and lime-mudstones in the lowest part of the Tober Colleen Formation indicates tranquil sedimentation below wavebase, in a basinal environment mostly free of bioturbation (Strogen, Jones and Somerville, 1990) |
| Waulsortian limestone Fm | Typically comprises pale-grey and very fine-grained (calcilutite-grade) carbonates, which display mudstone to wackestone depositional textures (Murray and Henry, 2018) | During the early Tournaisian, a major marine transgression inundated the landmass of Ireland and was followed by a period of carbonate ramp sedimentation (Murray and Henry, 2018) |

*Table 1 Detailed lithological descriptions and depositional modes of the rock formations located in the sampling area.*

### *2.1.1.2 Known unknown samples*

14 “known unknown” samples were added which originated from quarry source 6. They are described as “known unknowns” as their quarry of origin is known, however, they have not been included in the quarry classification process as these samples were acquired after the original geochemical analyses took place. These are thus used to test the success rate of the machine learning-based quarry classification scheme. As these samples originated from quarry source 6, they can be regarded as compositional and stratigraphic equivalents of any source 6 samples already included in the classification mechanism. Therefore, any sample information provided regarding Source 6 also relates to these 14 samples.

### 1.1.2 Pyrite textural information

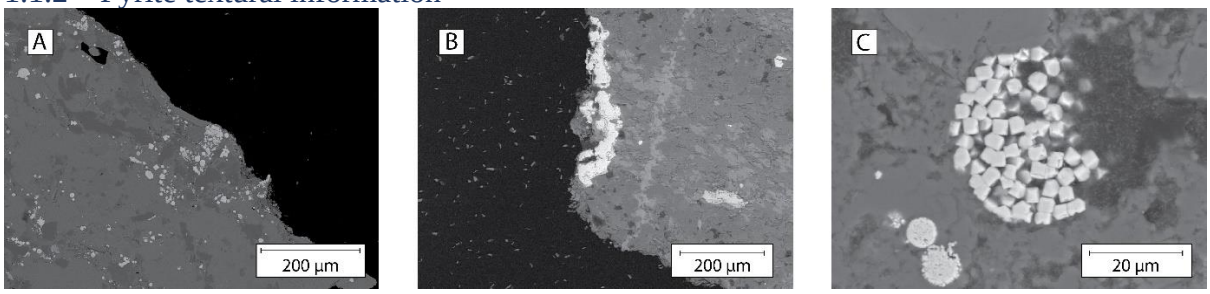



*Figure 2 (Double column, black and white) BSE (Back scattered electron) images of pyrite textural variation within the quarry samples. (A) Carbonate rock fragment containing framboidal and idiomorphic pyrite. (B) A cluster of anhedral idiomorphic pyrite. (C) Collection of framboidal pyrites displaying a range of microcrystal packing structures.*

Reflected light petrography and SEM analysis revealed two main pyrite morphologies within the quarry material: idiomorphic pyrite and framboidal pyrite. Each of these crystal morphologies are highly variable in shape and size (Figure 2) but are both commonly found within all six quarry samples. The idiomorphic pyrite grains vary in shape from anhedral to euhedral, while the microcrystals which form the framboidal pyrite grains, are highly variable in their packing structure. Individual framboids may appear very loosely packed together with clear space between the microcrystals, others are more tightly packed with minimal space in between the microcrystals. Additional BSE images of the quarry material, along with further information such as size ranges, can be found in Dornan *et al*, 2019.

### 2.1.3 Sample preparation

The sample material was first crushed and then subsequently sieved to retain the 250 µm – 4 mm particle size to provide a representative sample. The sieved material was then mounted in epoxy resin pucks. Resin mounted thin sections were then made from slices of these pucks. For, SEM – EDS analysis, resin mounted thin sections were exclusively used, while a combination of both polished pucks and resin mounted thin sections were used during the LA-Q-ICPMS analysis. Multiple replicates of the same samples were analysed throughout the analyses. These replicates were easily created by either re-polishing the surface of the resin mounted pucks, or creating a fresh resin mounted thin section from slices of the epoxy puck. For IRMS analysis, these crushed and sieved rock fragments were finely powdered using a TEMA mill and then, subsequently, decarbonated using 3 step acid digestion using 1 M HCl.

## 2.2 SEM – EDS

In-situ major element analysis of the pyrite crystals was undertaken using a Tescan MIRA XMU field emission scanning electron microscope (FE-SEM) equipped with an Oxford X-max 80mm$^2$ Energy Dispersive Spectrometer at the Centre for Microscopy and Analysis (CMA)/iCRAG Lab in Trinity College Dublin. In order to reach a dead time of ~30% we choose beam conditions 20 kV and 200 pA, with a working distance of 15-18.5 mm and a counting time of 30 seconds using both natural pyrite (Fe, S) and pure metal standards (Co, Ni, As, In). Detection limits on the SEM-EDS system are primarily related to the acquisition time (total number of counts per spectra) (Newbury and Ritchie, 2015). Routine quantitative analysis with the conditions as described in the current work give detection of approximately 0.1 wt %. All pyrites analysed during the SEM – EDS were confirmed using the pyrite structural formula Dornan *et al*, 2019)

## 2.3 LA – Q – ICPMS

LA-Q-ICPMS analysis were carried out at in the iCRAG Raw Materials Characterisation Laboratory in Trinity College Dublin, using a 193 nm Teledyne CETAC Analyte G2 ArF excimer laser coupled to a Thermo Fisher Scientific iCAP-Qc. Ablation occurs within a HelEx II two-volume ablation cell, using He carrier gas (c. 0.5 l/min) and a small volume of high-purity signal-boosting $N_2$ (c. 8 ml/min). Ar nebuliser gas (c. 0.55 l/min) is added to the line just before introduction to the mass spectrometer. An in-house adjustable-volume signal smoothing device was used to obtain a steady signal.

Tables 2 and 3 illustrate the laser parameters and analyte list used throughout the LA-Q-ICPMS analyses. Multiple test runs were carried out in the early stages of the analyses to investigate which trace elements were present in the pyrite samples. Elements such as Au, V, In and Tl were all included in these early analysis stages, however, Au, V and In were all found to be below the detection limits of the instrument, while Tl and Ni were found to have interferences with $^{208}Pb$ and $^{58}Fe$ respectively. As a result, none of the elements were included in later analyses. Regular ablations of the carbonate matrix were also conducted to verify that no contamination of the pyrite trace element content was occurring due to contribution from carbonate matrix.

The analytical procedure utilised a sample - standard bracketing procedure with blocks of reference materials separated spot analyses of pyrite. The USGS polymetal sulphide standard MASS-1 (Wilson *et al*, 2002) was used as the primary calibration standard and $^{57}Fe$ as the internal standard, with Fe concentrations for each quarry source taken from SEM-EDS analyses. MUL-ZnS-1 (Onuk *et al.*, 2017) and BCR-2G were used as secondary quality control standards to check the analytical accuracy. The preferred trace element concentrations for each of the reference materials are from the GeoReM database (http://georem.mpch-mainz.gwdg.de/). Mean measured values and accuracy to referenced values for each analyte are listed in table 4. Data reduction and production of trace element concentrations were undertaken using Iolite v3 using the trace element data reduction scheme (Paton *et al.*, 2011).

| Analyte | Dwell time (s) |
|---|---|
| $^{34}S$ | 0.01 |
| $^{43}Ca$ | 0.01 |
| $^{57}Fe$ | 0.01 |
| $^{59}Co$ | 0.03 |
| $^{63}Cu$ | 0.01 |
| $^{67}Zn$ | 0.06 |
| $^{75}As$ | 0.04 |
| $^{77}Se$ | 0.1 |
| $^{95}Mo$ | 0.1 |
| $^{107}Ag$ | 0.1 |
| $^{121}Sb$ | 0.1 |
| $^{208}Pb$ | 0.1 |

*Table 2 Analyte list used during the LA-Q-ICPMS analysis. Also included is the dwell time used per analyte*

| Fluence | Rep rate | Shot count | Spot size |
|---|---|---|---|
| 0.75-0.84 J/cm$^2$ | 5-6 Hz | 120-148 | 15 µm x 15 µm |

*Table 3 Laser parameters used during the LA-Q-ICPMS analysis*

| Element | Measured value (ppm; mean; n=3) | | | Referenced value (ppm) | | | Accuracy (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | MUL | MASS | BCR | MUL | MASS | BCR | MUL | MASS | BCR |
| Co | 348.67 | 71.00 | 36.50 | 308.00 | 60.00 | 38.00 | 13.20 | 18.33 | -3.95 |
| Cu | 1008.00 | 166000.00 | 16.93 | 994.00 | 134000.00 | 21.00 | 1.41 | 23.88 | -19.40 |
| Zn | 617000.00 | 219666.67 | 181.75 | 585500.00 | 210000.00 | 125.00 | 5.38 | 4.60 | 45.40 |
| Se | 66.23 | 73.80 | 13.55 | 66.00 | 65.00 | - | 0.35 | 13.54 | - |
| As | 207.67 | 46.87 | 1.18 | 200.00 | 51.00 | - | 3.83 | -8.10 | - |
| Mo | 59.60 | 49.53 | 207.93 | 60.00 | 59.00 | 270.00 | -0.67 | -16.05 | -22.99 |
| Ag | 620.67 | 46.10 | 0.61 | 607.00 | 50.00 | 0.50 | 2.25 | -7.80 | 22.50 |
| Sb | 820.67 | 65.70 | 0.46 | 819.00 | 60.00 | 0.35 | 0.20 | 9.50 | 31.43 |
| Pb | 1197.33 | 68.77 | 7.46 | 1149.00 | 68.00 | 11.00 | 4.21 | 1.13 | -32.18 |

*Table 4 Mean measured values for reference materials and accuracy of measured values to referenced values.*

## 2.4 IRMS

Isotope analyses were carried out at the University of St. Andrews with an EA Isolink coupled to a MAT253 IRMS via a Conflo IV (Thermo Fisher Scientific, Bremen, Germany). Decarbonated rock powders were weighed into 8 x 5 mm tin capsules and combusted at 1020°C under a constant He stream (flow rate 100 ml/min, dropping to 50 ml/min after 20 seconds) with a 5-second pulse of $O_2$ gas (flow rate of 250 ml/min) to convert all sulphide to $SO_2$ gas. Tungstic oxide granules were used as an additional combustion aid in the reactor. The tungstic oxide was followed by copper wires to reduce the minor $SO_3$ to $SO_2$. Water resulting from the combustion was trapped at room temperature in a separate column packed with magnesium perchlorate grains. The remaining $SO_2$ was further purified with a GC column at 45°C. The international reference standards IAEA-S2 and IAEA-S3 were included at the beginning and end of each run for calibration. Analytical accuracy was monitored with IAEA-S1, which agreed with internationally recognized values to within < 0.5‰. Peak areas were calibrated for abundance measurements with of a series of sulphanilamide standards. The isotopic data are expressed in standard delta notation relative to VCDT.

## 2.5 Machine learning

Machine learning methods employ computational algorithms that attempt to emulate the process of human intelligence and neural networks by learning from data fed into the system. These algorithms and models are integral to “big data” analysis (El Naqa and Murphy, 2015). Machine learning methods are prevalent in many aspects of geosciences such as hazard modelling (Yilmaz, 2009; Wang, Sawada and Moriguchi, 2013) and mineral prospecting (Carranza and Laborte, 2015; Xiong and Zuo, 2018), however, they are also becoming increasingly applied to large geochemical datasets(e.g. Rodriguez-Galiano *et al.*, 2012; Gregory *et al.*, 2019).

Two main types of machine learning software packages were used: The Aggregate Quarry Classification Model (AQCM) and Waikato Environment for Knowledge Analysis software (WEKA). The AQCM machine learning software was created for this project and applied to the data generated by geochemical analysis of bulk rock aggregate and pyrite described in this paper. This software package has been made open source and can be used for the classification of any geochemical database. Further information regarding how to operate the AQCM and its potential uses can be found using the GitHub link attached to this paper.

WEKA is a free to download collection of machine learning algorithms for data mining techniques. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization (Witten *et al.*, 2016). This software package is coded in the Java programming language and allows users to harness machine learning models without any prior computer programming knowledge
Both software packages use logistic regression and random forest learning models in order to classify and predict the quarry sources of aggregate fill based on their geochemical composition. Two different pieces of software were used in order to compare the results of the AQCM with that of a freely available piece of software.

### 2.5.1 Logistic regression model

Logistic regression is a mathematical modelling approach that can be used to characterise the relationship between a number of variables (e.g. major and trace element composition, $\delta^{34}S$ signature) to a multichotomous dependant variable (e.g. quarry sources 1 – 6). Logistic Regression is a statistical model, based on the logistic function (Equation 1), which relates the probability of a given event to a linear combination of predictors (Equation 2). The parameters of the linear equation ($\alpha$ and $\beta_1 \ldots \beta_n$) are then learned from the data such that they maximise the model's prediction accuracy (Kleinbaum and Klein, 2002; Yilmaz, 2009). The logistic regression model is useful for predicting the outcome of a multivariate analysis based on values of a set of predictor variables (Yilmaz, 2009). As a result, it can be used to investigate the likelihood of a given pyrite/ bulk rock analysis to be identified as one of the quarry sources 1-6.

Equation 1: $f(z) = 1 / (1+ e^{-z})$
Equation 2: $z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$

### 2.5.2 Random forests

Random forests are a supervised method of data classification constructed from a number of decision trees (Breiman, 2001). These decision trees follow a flowchart-like structure, in which splits, or nodes, partition the dataset using a random subset of input variables (e.g. < 500 ppm Co or > 350 ppm Pb). This unbiased selection increases the diversity and robustness of the random forest model as each decision tree classifies samples based on a unique series of random tests (Rodriguez-Galiano *et al.*, 2012). Partitioning of the dataset starts with the "Root node" which integrates the entire data set before it is split into two groupings. Splitting occurs until a "leaf node" is reached. "Leaf nodes" represent a discrete class label (e.g. quarry 1-6) and indicate that the sample has been classified based on the characteristics of a predefined grouping. Each decision tree within a random forest contributes a single vote for the assignment of the most frequent class to the input data (Breiman, 2001; Rodriguez-Galiano *et al.*, 2012) . As a result, sample classification by random forests is based on the modal classification of several decision trees.

## 2.6 Statistical analysis

All statistical analyses were carried out using IMDEX ioGAS 7.0 advanced geochemical data analysis software.

### 2.6.1 Principle component analysis (PCA)

PCA is an unsupervised statistical method that rotates and shears a data matrix of *n*-dimensions along axes (components) of greatest variability (Hammer, 2017). Unsupervised statistical transformations are not influenced by categorical data, and the transformation of a dataset will always be the same using PCA given the same numeric data matrix. The number of components in a dataset is equal to the number of input variables, or one less than the number of datapoints, whichever is smaller.

Components are ranked, such that component 1 is the axis of greatest variance in a dataset, component 2 is orthogonal to component 1 and so on. The usefulness of PCA as a data interpretation tool stems from the fact that PCA often permits dimension reduction of multivariate data by the discarding of the lower ranked components of often low variance, thus permitting the interpretation of multidimensional data, previously existing in a hyperspace, in a 2- or 3-dimensional space (e.g. on biplot). An advantage of PCA versus two- or three-element or element-ratio biplots and triplots is that PCA integrates information from *n* variables, and thus can potentially act as a much more powerful discrimination or plotting tool than biplots and triplots that are limited in the amount of information that they integrate.

For PCA, if the ppm values of the various trace, minor and major elements are examined in isolation, and not in the context of the entire ablated volume, misleading determinations may result (Pawlowsky-Glahn, V. and Buccianti, 2011). As a result, the data used in PCA in this paper has been transformed using centred log ratio (CLR) transformation, including a residual value representing non-analysed elements to sum to unity (i.e. 1 million ppm). CLR is calculated as the log of the individual measurement divided by the geometric mean of that element across the entire dataset. CLR transformation can be quickly calculated using software such as ioGAS, but also by freeware such as CoDaPack or "R".

# 3. Results

## 3.1 Geochemical Results

The results obtained from the SEM-EDS, LA – Q – ICPMS and IRMS analysis of quarry samples 1-6 are presented in table 5 and summarised in 2-D plots 3 - 5. For LA -Q – ICPMS analysis, where pyrite grains had trace element concentrations below the detection limits for the analytes, their concentrations were substituted for half the minimum limit of detection for the run in which they were analysed . Nonparametric statistics are used due to the non – normal distribution of the dataset.

| No. of analyses | Analyte | Quarry 1 | Quarry 2 | Quarry 3 | Quarry 4 | Quarry 5 | Quarry 6 |
|---|---|---|---|---|---|---|---|
| 69 | TS (%) | 1.87 | 0.65 | 0.04 | 0.64 | 1.02 | 1.33 |
| | $\delta^{34}S$ (%) | -26.95 | -7.91 | 4.98 | -1.98 | -24.29 | -20.72 |
| 274 | Fe (wt%) | 43.53 | 43.60 | 39.18 | 43.30 | 43.08 | 43.17 |
| | S (wt %) | 51.80 | 51.61 | 44.28 | 49.37 | 50.86 | 49.64 |
| 491 | Co (ppm) | 8.70 | 62.90 | 230.00 | 42.00 | 187.00 | 58.00 |
| | Cu (ppm) | 45.00 | 560.00 | 233.00 | 73.00 | 305.00 | 79.00 |
| | Zn (ppm) | 4.31 | 104.00 | 59.46 | 7.92 | 156.50 | 7.29 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | As (ppm) | 78.00 | 246.00 | 773.00 | 2390.00 | 417.50 | 35.20 |
| | Se (ppm) | 242.00 | 77.00 | 4.69 | 6.10 | 44.95 | 215.00 |
| | Mo (ppm) | 12.10 | 67.40 | 18.80 | 14.00 | 30.50 | 16.60 |
| | Ag (ppm) | 0.08 | 7.60 | 2.60 | 2.10 | 2.85 | 0.09 |
| | Sb (ppm) | 7.50 | 22.60 | 123.00 | 88.00 | 28.45 | 5.60 |
| | Pb (ppm) | 10.10 | 278 | 340.00 | 530.00 | 152.50 | 14.70 |

*Table 5 Median concentrations of quarry sources 1 – 6 from the geochemical analysis of pyrite and bulk rock aggregate.*

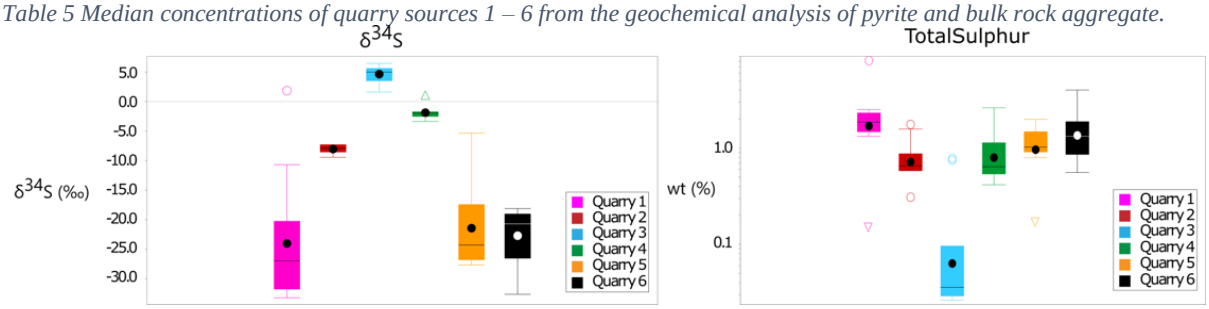



*Figure 3 (Double column, colour) Box and whisker plot of IRMS analysis of bulk rock aggregate from quarry sources 1 – 6.*

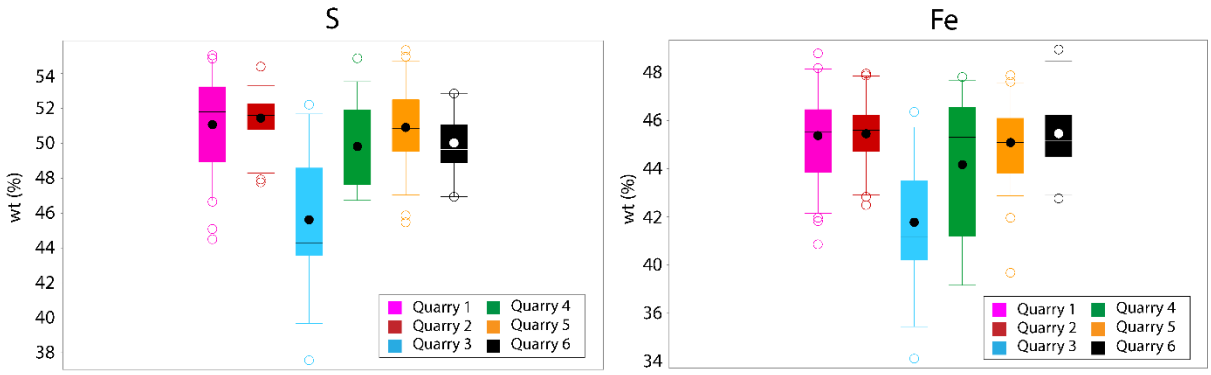



*Figure 4 (Double column, colour) Box and whisker plot of SEM - EDS analysis of pyrite from quarry sources 1 – 6.*

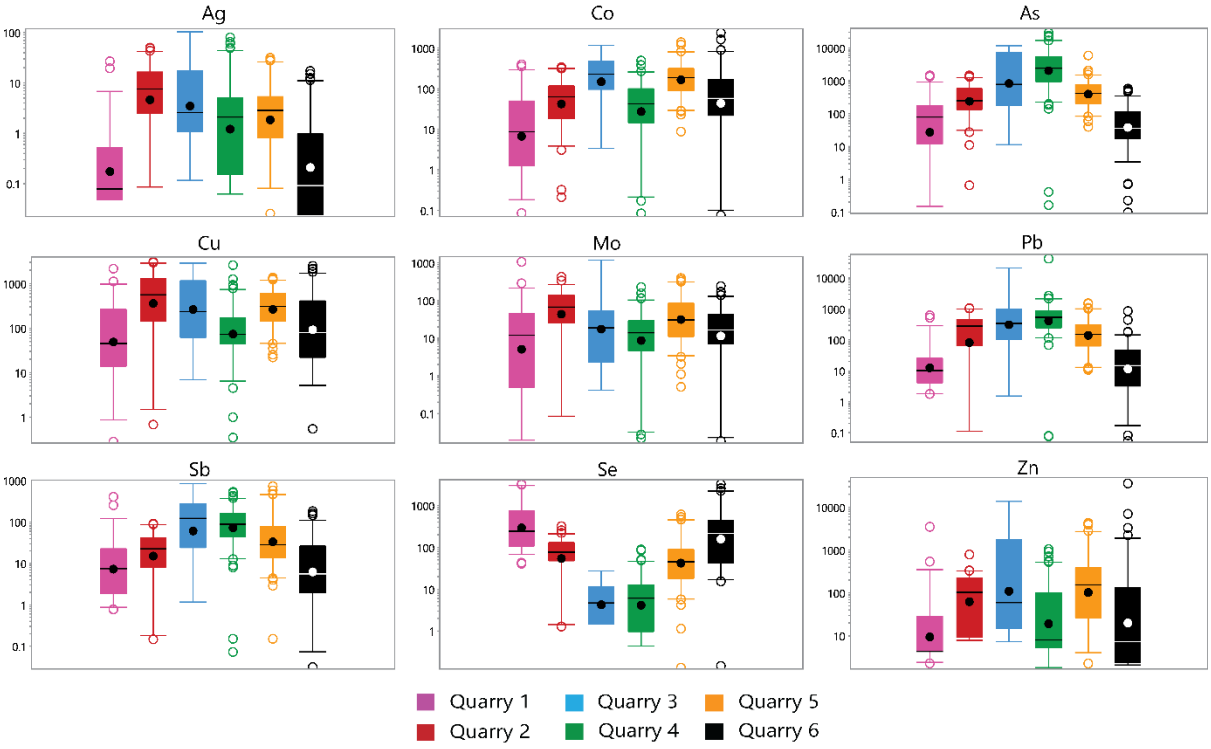



*Figure 5 (Double column, colour) Box and whisker plot of LA - Q - ICPMS analysis of pyrite from quarry sources 1 – 6.*

## 3.2 PCA results

The results from the geochemical analysis of pyrite and bulk rock aggregate were analysed by PCA to enhance the geochemical variance between quarry sources 1 – 6. This enhancement in the variance was hoped to improve source classification when using the machine learning models

CLR transformation was applied to the data obtained through the geochemical analysis of pyrite and bulk rock aggregate. This CLR transformed data was then analysed by PCA. Table 6 illustrates the factor loadings for principle components (PCs) 1 – 7 of the SEM – EDS, LA – Q – ICPMS and IRMS combined dataset, these PCs account for 90 % of the variance in the dataset. This data is also summarised in Figure 6.

In addition to this combined dataset, CLR transformation and PCA were also performed on individual datasets (e.g. Fe, S concentrations from SEM – EDS analysis) along with different combinations of datasets (e.g. major and trace element concentrations). All combinations of data evaluated by PCA are outlined in Tables 5 and 7. However, results for the PCA analysis of these datasets are not presented in this paper but can be found in the supplemental material.

| Eigenvectors | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
|---|---|---|---|---|---|---|---|
| Co | 0.02 | 0.17 | 0.80 | 0.25 | 0.08 | 0.14 | 0.33 |
| Cu | 0.15 | 0.30 | -0.07 | -0.05 | 0.70 | -0.46 | -0.05 |
| Zn | 0.15 | 0.25 | -0.26 | 0.61 | -0.15 | 0.38 | -0.19 |
| As | 0.19 | -0.42 | 0.13 | -0.15 | 0.00 | 0.03 | 0.05 |
| Se | -0.19 | 0.44 | -0.22 | -0.27 | 0.04 | 0.10 | 0.04 |
| Mo | 0.17 | 0.28 | 0.20 | -0.43 | -0.55 | -0.20 | -0.37 |
| Ag | 0.25 | -0.03 | -0.33 | 0.29 | -0.28 | -0.44 | 0.44 |
| Sb | 0.23 | -0.27 | -0.18 | -0.38 | 0.04 | 0.30 | 0.46 |
| Pb | 0.23 | -0.31 | -0.05 | 0.01 | 0.28 | 0.26 | -0.50 |
| Stoichiometry | -0.41 | -0.10 | -0.05 | 0.00 | -0.01 | -0.01 | 0.01 |
| TS (%) | -0.39 | 0.03 | -0.07 | -0.03 | 0.03 | 0.15 | 0.10 |
| δ34S (%) | -0.18 | -0.41 | 0.12 | 0.24 | -0.09 | -0.45 | -0.21 |
| Fe (wt%) | -0.41 | -0.10 | -0.05 | 0.00 | -0.01 | -0.01 | 0.00 |
| S (wt %) | -0.41 | -0.09 | -0.06 | 0.00 | -0.01 | -0.01 | 0.01 |
| Variance % | 40.62 | 17.97 | 8.41 | 7.49 | 6.24 | 5.43 | 4.04 |
| Cumulative variance % | 40.62 | 58.59 | 67.00 | 74.49 | 80.73 | 86.16 | 90.19 |

*Table 6 Results from the PCA of the entire geochemical dataset (IRMS, SEM – EDS and LA – Q – ICPMS)*

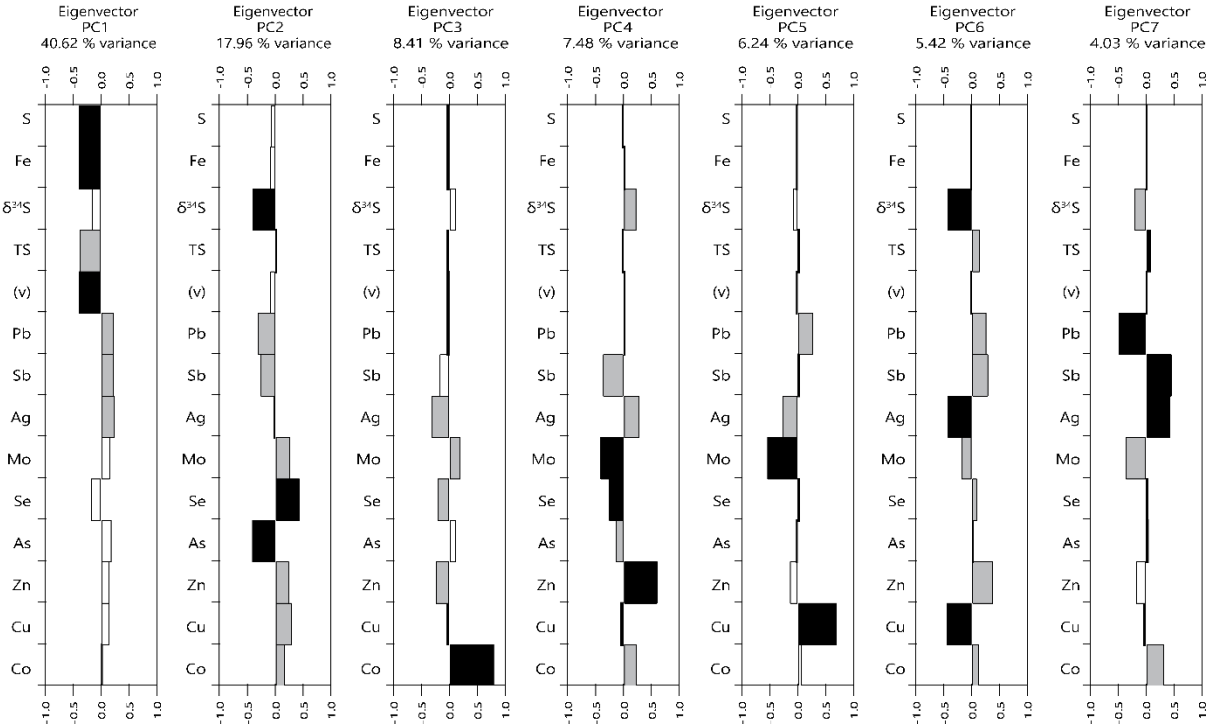



*Figure 6 (Double column, black and white) Factor loadings plot for PCs 1 – 7. Analytes with a factor loading score of ≥ 0.4 are shaded black, between 0.2 – 0.4 shaded grey and ≤ 0.2 are shaded white. (v) represents the stoichiometry.*

## 3.3 Machine learning classification

The raw geochemical data obtained from the SEM – EDS, LA – Q – ICPMS and IRMS analysis, as well as the PCA transformed data, were classified using random forest and logistic regression models. The results obtained from the classification of quarry sources 1 – 6 by logistic regression and random forest are presented in tables 7 and 8. These results represent the accuracy with which the models can classify samples based on SEM – EDS, LA – Q – ICPMS and IRMS analyses. The accuracy scores for both models are based on 10-fold cross validation. *K*-fold cross validation is used ahead of the conventional train – test approach as it reduces model over fitting, while also producing more reliable and unbiased testing compared to the train - test approach. Table 9 illustrate the mean ROC area, F-measure and Kappa statistic for the machine learning models. These values indicate the prediction capability of the classification technique and also illustrate how optimal the models are for classifying the geochemical data. Figure 7 is an illustration of the tests carried out within a decision tree in order to classify samples in the WEKA random forest model.

| Dataset | AQCM | | WEKA | |
|---|---|---|---|---|
| | Random forest | | Random forest | |
| | Raw | PCA | Raw | PCA |
| Majors | 62.37% | 51.26% | 61.31% | 58.03% |
| Traces | 55.73% | 56.87% | 85.35% | 78.95% |
| S isotopes | 62.56% | 43.43% | 63.49% | 47.62% |
| Majors; traces | 97.06% | 69.40% | 100% | 90.68% |
| Majors; S isotopes | 100% | 98.24% | 100% | 99.64% |
| S isotopes; traces | 98.14% | 73.40% | 100% | 95.42% |
| Majors; traces; S isotopes | 98.83% | 80.72% | 100% | 96.79% |

*Table 7 Results for the classification of quarry sources 1 – 6 by AQCM and WEKA using random forest classification model*

| Dataset | AQCM | | WEKA | |
|---|---|---|---|---|
| | Logistic regression | | Logistic regression | |
| | Raw | PCA | Raw | PCA |
| Majors | 31.44% | 61.75% | 31.75% | 24.81% |
| Traces | 60.85% | 63.62% | 73.22% | 73.22% |
| S isotopes | 54.52% | 29.70% | 69.85% | 38.09% |
| Majors; traces | 93.23% | 90.80% | 99.54% | 98.16% |
| Majors; S isotopes | 90.74% | 81.38% | 99.64% | 100% |
| S isotopes; traces | 95.88% | 93.26% | 99.38% | 98.39% |
| Majors; traces; S isotopes | 99.56% | 97.54% | 100% | 98.17% |

*Table 8 Results for the classification of quarry sources 1 – 6 by AQCM and WEKA using logistic regression classification model*

| Dataset | Prediction capability | | | | | |
|---|---|---|---|---|---|---|
| | Logistic regression | | | Random forest | | |
| | ROC Area | F-Measure | Kappa Statistic | ROC Area | F-Measure | Kappa Statistic |
| Majors | 0.65 | 0.28 | 0.15 | 0.82 | 0.60 | 0.53 |
| Traces | 0.94 | 0.72 | 0.66 | 0.97 | 0.85 | 0.82 |
| S isotopes | 0.87 | 0.68 | 0.64 | 0.89 | 0.63 | 0.56 |
| Majors; traces | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 |
| Majors; S isotopes | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 |
| S isotopes; traces | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 |
| Majors; traces; S isotopes | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

*Table 9: Mean values for ROC, F-Measure and Kappa statistic for both logistic regression and random forest models. Receiver operating characteristic (ROC) measurement values illustrate model optimisation. Values approaching 1 indicate the model is optimised for data classification, while values approaching 0.5 are comparable to random guessing and a non-optimal model choice. F-Measure is a combined measure of precision and recall (2 × recall × precision)/(recall+precision). Kappa statistic is a chance corrected error (success rate of actual predictor - success rate of random predictor) / (1 - success rate of random predictor)* (Witten *et al.*, 2016)

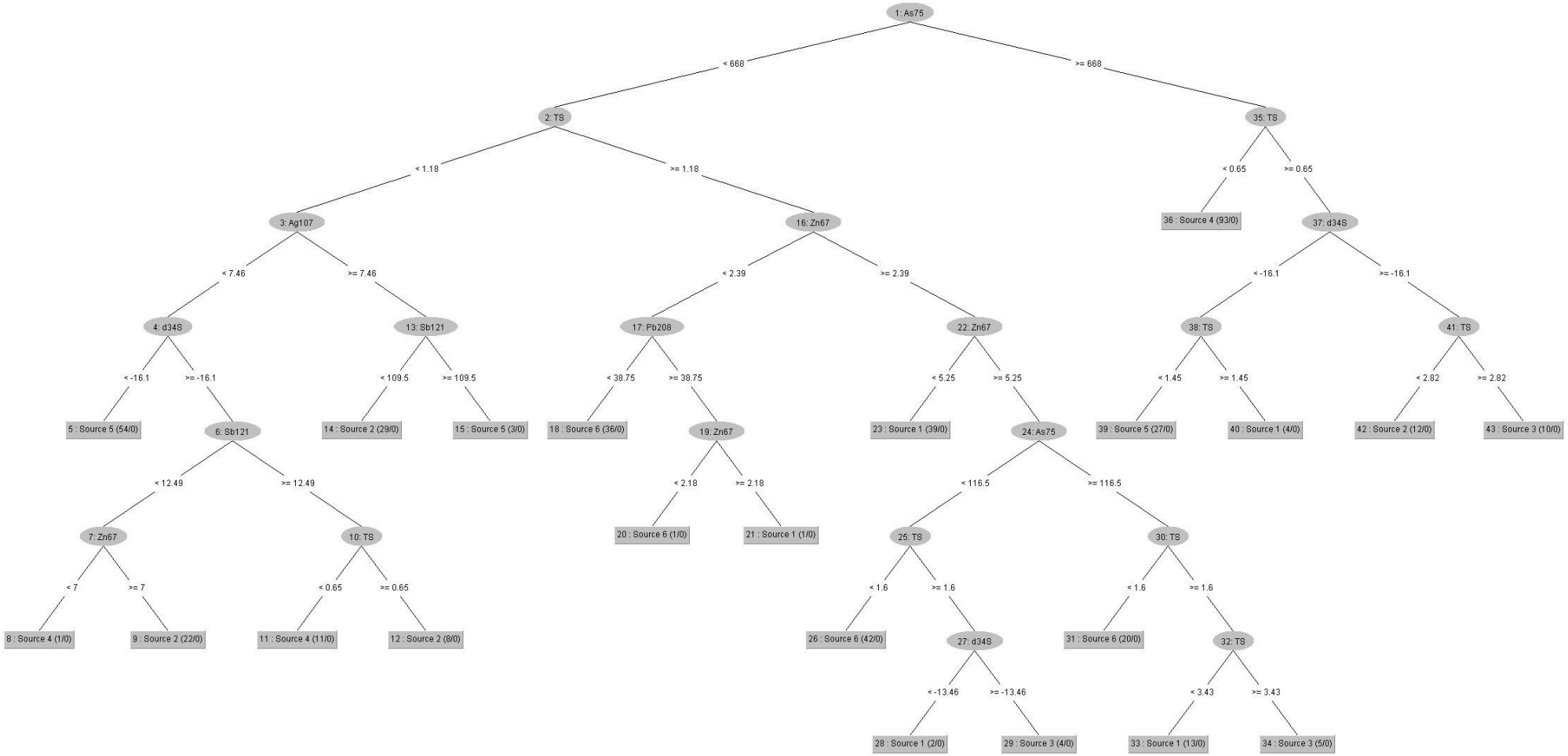



*Figure 7 (Full page, black and white) Example of decision tree used as part of the random forest classification of quarry sources 1 - 6 using WEKA software.*

### 3.3.1 Classification of known unknown samples

As the LA – Q – ICPMS dataset achieved the highest average classification score amongst the non-combined datasets (Table 7 and 8), it was selected to classify a set of known unknown samples through a train-test classification mechanism. This known unknown dataset comprised of 309 spot analyses of pyrite. Ten randomly selected sub samples of 50 analyses were extracted from this dataset and used as "test" suite to investigate the capabilities of the machine learning models to classify unknown samples. The results obtained from the classification of these known unknown samples by Random Forest and Logistic regression are outlined in table 10.

| Sub sample | AQCM | | WEKA | |
|---|---|---|---|---|
| | Logistic regression | Random forest | Logistic regression | Random forest |
| 1 | 64 | 50 | 78 | 46 |
| 2 | 70 | 62 | 74 | 34 |
| 3 | 64 | 54 | 70 | 36 |
| 4 | 54 | 40 | 64 | 34 |
| 5 | 56 | 28 | 66 | 30 |
| 6 | 60 | 40 | 72 | 38 |
| 7 | 62 | 44 | 74 | 28 |
| 8 | 60 | 44 | 62 | 28 |
| 9 | 68 | 44 | 78 | 46 |
| 10 | 70 | 42 | 70 | 38 |

*Table 10 Results for the classification of 10 random sub samples of known unknown LA – Q – ICPMS dataset. Results indicate the percentage of the 50 known unknown samples that were classified correctly.*

# 4. Discussion

## 4.1 Geochemical analyses and quarry source discrimination

Analysis of pyrite and bulk rock aggregate from quarry sources 1 – 6 reveals compositional variability between the different sources. Each of the analytical techniques vary in their viability to discriminate each of the quarry sources. These compositional differences are least apparent when examining pyrite major element geochemistry (Figure 8). From experimental analysis, stoichiometric pyrite contains 53.5 % S and 46.5 % Fe (Anthony *et al.*, 1995), as a result, a significant amount of compositional overlap exists between quarries 1, 2, 4, 5 and 6 surrounding these concentrations. However, quarry 3, which has median concentrations of 39.18 % Fe and 44.28 % S, is easily distinguishable due to these low Fe and S concentrations. As described in Dornan *et al* (2020), factors affecting these low Fe and S values are the partial oxidation of these pyrites within the sample material the presence of trace elements which are present below the detection limits of the SEM technique. As the purpose of this study is to discriminate sources based upon all/any characteristics and not to analyse only unaltered products, the fact that these pyrites are oxidised is not detritmental to their characterisation, and in fact provides a very clear discriminatory criterion. However, it must also be noted that pyrite does not oxidise at a constant rate. Careful consideration must be applied when analysing historical material as misclassification of material may occur if "amount of oxidation" is used as a classification criteria.

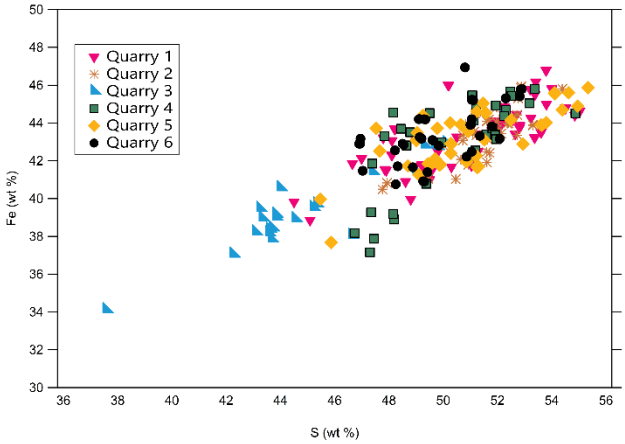



*Figure 8 (Single column, colour) Bi plot of Fe and S concentrations of quarry sources 1 – 6.*

By investigating the bulk rock geochemistry using IRMS, it is possible to reduce the number of quarry sources which overlap compositionally, due to the greater variation in TS and $\delta^{34}S$ concentrations. For example, the maximum variation in median Fe and S concentrations is 4.42 wt % and 7.52 wt % respectively, in contrast, $\delta^{34}S$ concentrations deviate by up to 26.91 %. Consequently, a simple bi plot of TS vs $\delta^{34}S$ illustrates a much clearer compositional separation of quarries 2, 3 and 4 from the remaining sources. Therefore, bulk rock S isotope geochemistry offers an improved method of quarry separation when compared to the major element geochemistry of pyrite.

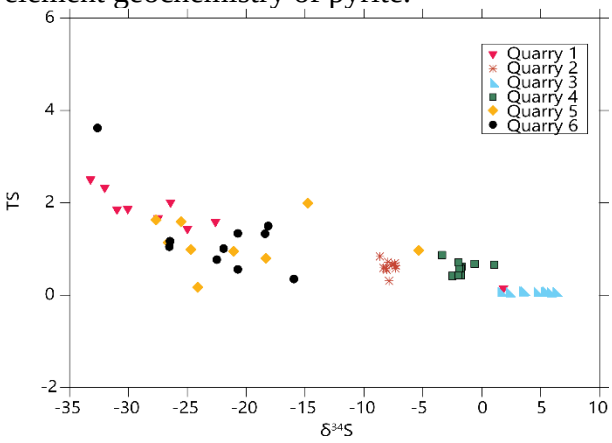



*Figure 9 (Single column, colour) Bi plot of TS vs $\delta^{34}S$ concentrations for quarry sources 1 – 6.*

With that being said, the analysis of pyrite trace element geochemistry by LA – Q – ICPMS provides the most dimensions with which to discriminate quarries. The degree with which trace elements are incorporated into pyrite can vary depending on the incorporation mechanism, type of pyrite precursor mineral present and local oxidation conditions (Dellwig *et al.*, 2002). As a result, median concentrations for elements such as As, Se and Pb can vary between the quarry sources by up to 1897 ppm, 302 ppm and 400 ppm respectively. When these trace element concentrations are plotted on bi-plots or ternary diagrams, individual quarry sources can be visually distinguished (Figures 10 and 11). This variance is enhanced through the use of PCA. Figure 12 is a plot of PC1 and PC2 for the trace element dataset, the arrow length for each element in this plot indicates how well the element explains the variance in the dataset, with a longer arrow length indicating a stronger influence on the variance. Consequently, elements such as As, Cu, Pb and Se, all of which have long arrow lengths, are identified as elements which strongly influence the variance in the dataset.

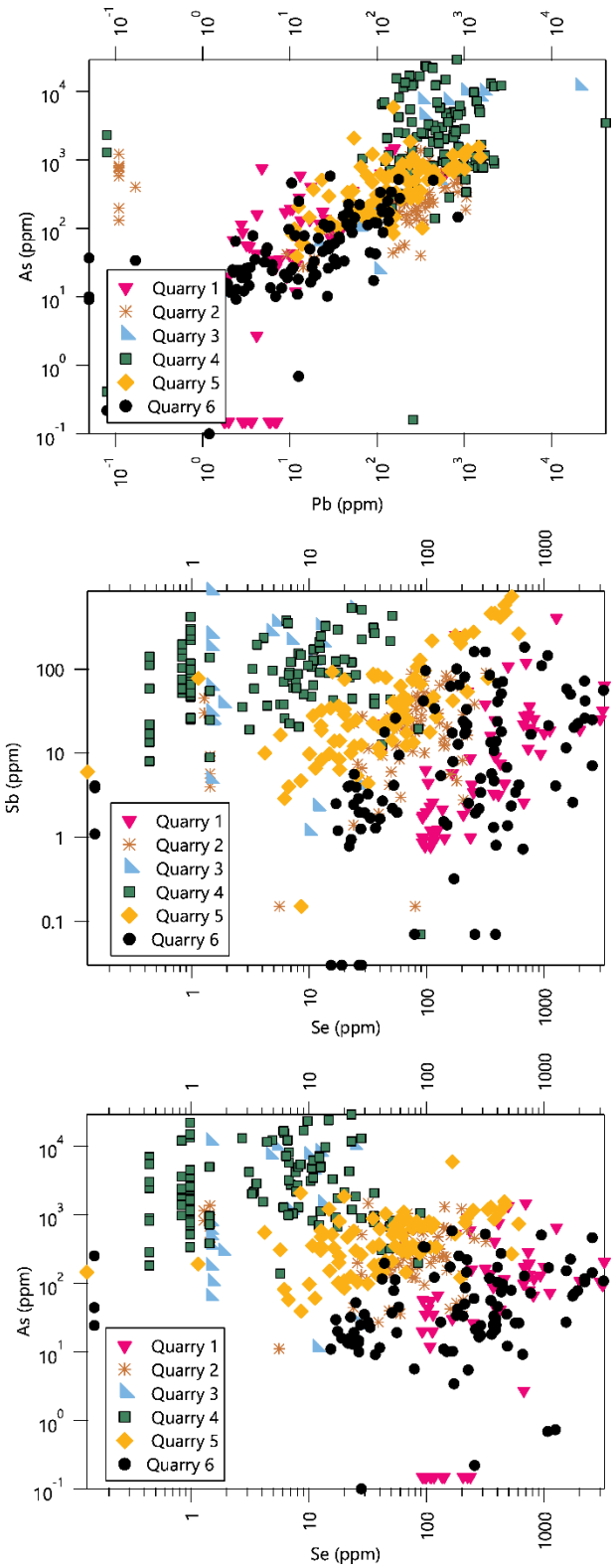



*Figure 10 (Single column, colour) Bi plots of trace element concentrations for quarry sources 1 – 6.*

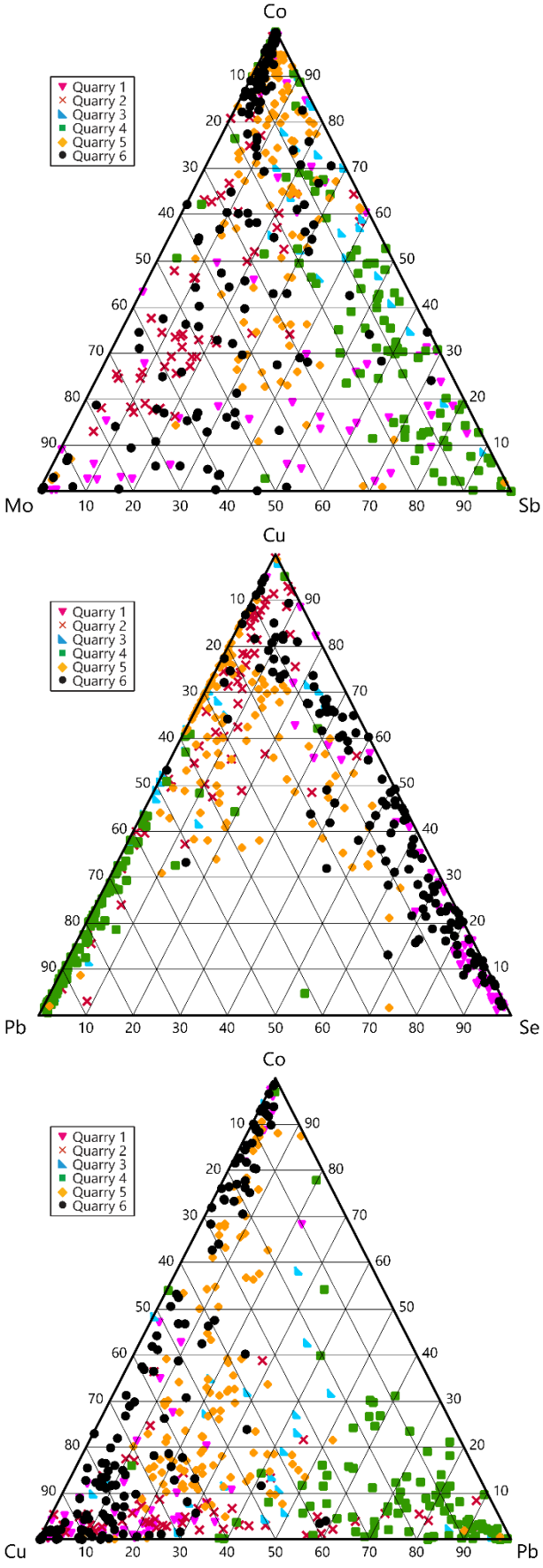
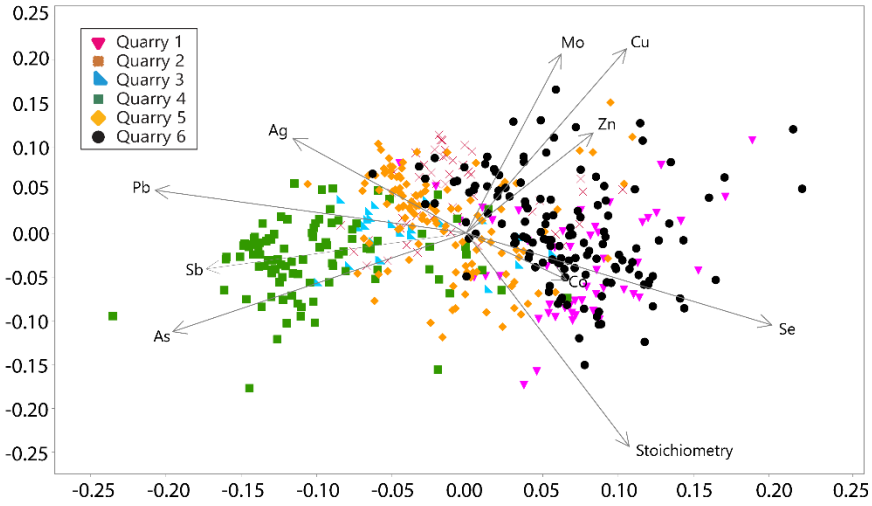



*Figure 11 (Single column, colour) Ternary plots of trace element concentrations for quarry sources 1 – 6.*




Despite some compositional separation being offered by each of these analytical techniques,

*Figure 12 PCA plot of PC1 vs PC2 for trace element concentrations of quarry sources 1 – 6.*

the classification provided by clustering and other classifier methods was unsatisfactory. In Dornan *et al* (2020), k-means clustering was used to quantify the compositional separation using Fe and S concentrations. This was similarly attempted using the trace element and S isotope datasets; however, the results were inadequate due to the constant compositional overlap between certain sources. Therefore, classification of the quarry sources based on their geochemical composition was not possible using clustering techniques.

## 4.2 Machine learning and quarry classification

In Dornan *et al*, (2019) k-means clustering was used a method of quarry source classification. However, this method proved unsuccessful in accurately classifying quarry sources 1 - 6, therefore, an improved method of quarry classification was needed. This led to the use of machine learning models which provided a much more powerful and effective method of classification. Two different machine learning models were used as distinctive information could be gleaned from the results of both models. For instance, as the logistic regression model is designed to describe probability, the results from this model give the likelihood of a sample being classified from each of the six quarry sources. This information can be found in the output of the logistic regression model when using the AQCM or WEKA.

Each model was tested using both PCA data and raw compositional data. PCA data was included as it enhances the visualisation of variance within a dataset. However, results of both models indicated that the inclusion of PCA caused a reduction in performance of both models, probably due to the dimension-reduction inherent to PCA. For example, using the random forest model, classification accuracy decreased by up to 27 %, while in the logistic regression model performance dropped by almost 32 %. As a result, only raw compositional data should be used for classification purposes.

As previously stated, both individual datasets and a combination of datasets were used as part of the classification process to investigate which dataset offered the best classification accuracy. For the individual datasets, the trace elements provided the greatest classification accuracy in the random forest model (77.32 %) and logistic regression models (63.10 % and 65.70 %). As result, this dataset was selected to classify of a set of known unknown samples. These “known unknowns” were part of a separate dataset for Source 6, unused in the original

classification database. Ten sub-samples from this dataset were used as test samples to investigate the model's capability to classify unknowns. In this experiment, the logistic regression model proved much more powerful at classifying unknowns compared to the random forest model. It achieved an average classification score of 66.8 % while the random forest model only achieved an average score of 40.3 %.

A simple way to increase the classification accuracies of machine learning models is to use combinations of geochemical datasets. For instance, when the trace element dataset is combined with S isotope dataset using the Logistic regression model, model accuracy increases by from 63.10 % to 95.49 %. This increase in accuracy is due to the use of median concentration values when combining datasets. Median values were chosen as the geochemical results of both SEM – EDS and IRMS analyses follow a non – normal distribution. When using the major element dataset in combination with the trace element dataset, median Fe and S concentrations for each quarry source are used. This is due to the extreme difficulty in linking the individual pyrites analysed by SEM – EDS with those analysed by LA – Q – ICPMS, as pyrites within these samples often measure ≤ 12 µm. Similarly, since a reduced number of samples were analysed by IRMS, median $\delta^{34}S$ and TS concentrations are applied when the S isotope dataset is used in combination with either major element or trace element datasets.

As a result, if this classification mechanism is used as part of a quarry classification system, statistically meaningful mean values would need to be used in in order to represent the geochemical composition of a quarry source. This would require a near-normally distributed dataset with a statistically meaningful number of samples through the quarry succession. However, due to the restricted access to quarry material this may not be a possibility.

# 5. Conclusions

- Using machine learning models, such as logistic regression and random forest, it is possible to generate a classification mechanism for aggregate quarry sources based on their bulk rock and pyrite geochemistry. Depending on the dataset used, these models can range in accuracy from 31 % to 100 %. However, when classifying known unknowns, the logistic regression model out preforms the random forest model by achieving an average classification score of 66.80 %.
- The accuracy of these models is enhanced using median concentration values when combining datasets. These median values are applied as the composition of the pyrites analysed by SEM – EDS and LA – Q- ICPMS follow a non – normal distribution. Additionally, relating pyrite crystals analysed by SEM – EDS to those analysed by LA – Q – ICPMS is extremely difficult, as pyrite crystals within these samples often measure ≤ 12 µm. Therefore, if this classification mechanism were used as part of an applied quarry classification system, statistically meaningful mean values taken from a near normally distributed dataset would have to be used in order to accurately represent the quarry composition.
- Although PCA was used as part of this investigation, it proved detrimental to the performance of the machine learning models. When using PCA data, the performance of the logistic regression and random forest models dropped by 32 % and 27 % respectively. This was counter intuitive to our original hypothesis as PCA was used to enhance the variance within a dataset. However, this enhancement in variance is achieved by rotating and shearing the dataset along orthogonal axes of greatest variability. This transformation reduced the dimensionality of the dataset and caused

the variables to appear monotonous to the machine learning models leading to a reduction in model performance.

## Acknowledgements

A massive thank you goes to Dr. Neal O'Riain for all his help and hard work in creating the AQCM. Without him most of this work would not have been possible. I would also like to thank Cora McKenna and Dr. Gary O'Sullivan for all their patience and time spent teaching me about machine learning and PCA through the years.

## Computer code availability

AQCM is available online as an open source Python script using the following link:
https://github.com/Tadhg-D/Aggregate-Quarry-Classification-Model-AQCM-

WEKA is available online using the following link:
https://www.cs.waikato.ac.nz/ml/weka/downloading.html

## Funding

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number 13/RC/2092 and co-funded under the European Regional Development Fund and by iCRAG industry partners.

## References

Anthony, John W, Bideaux, Richard A, Bladh Kenneth W and Nichols, M. C. (1995) 'Pyrite', in Anthony, John W, Bideaux, Richard A, Bladh Kenneth W and Nichols, M. C. (ed.) *Handbook of Mineralogy*. 1st edn. Chantilly, VA.: Mineralogical Society of America. doi: 10.1016/S1470-2045(10)70008-5.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Carranza, E. J. M. and Laborte, A. G. (2015) 'Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines)', *Computers and Geosciences*. Elsevier Ltd, 74, pp. 60–70. doi: 10.1016/j.cageo.2014.10.004.

Dellwig, O. *et al.* (2002) 'Trace metals in Holocene coastal peats and their relation to pyrite formation (NW Germany)', *Chemical Geology*. Elsevier, 182(2–4), pp. 423–442. doi: 10.1016/S0009-2541(01)00335-7.

Dornan, T., Goodhue, R. and Riegler, T. (2019) 'Discriminating aggregate sources with in-situ mineral chemistry: an Irish example', *Quarterly Journal of Engineering Geology and Hydrogeology*. doi: 10.1144/qjegh2018-176.

Gallagher, M. (2016) ‘Novel geochemical fingerprints of biogenicity applied to ancient carbonates’.

Geological Survey Ireland (2018) *Geological survey Ireland spatial resources*. Available at: https://dcenr.maps.arcgis.com/apps/MapSeries/index.html?appid=a30af518e87a4c0ab2fbde2aaac3c228 (Accessed: 15 July 2019).

Gregory, D. D. *et al.* (2015) ‘Trace Element Content of Sedimentary Pyrite in Black Shales *’, *Economic Geology*, 110, pp. 1389–1410. doi: 10.2113/econgeo.110.6.1389.

Gregory, D. D. *et al.* (2017) ‘Whole rock and discrete pyrite geochemistry as complementary tracers of ancient ocean chemistry: An example from the Neoproterozoic Doushantuo Formation, China’, *Geochimica et Cosmochimica Acta*. doi: 10.1016/j.gca.2017.05.042.

Gregory, D. D. *et al.* (2019) ‘Distinguishing Ore Deposit Type and Barren Sedimentary Pyrite Using Laser Ablation-Inductively Coupled Plasma-Mass Spectrometry Trace Element Data and Statistical Analysis of Large Data Sets’, *Economic Geology*, 114(4), pp. 771–786. doi: 10.5382/econgeo.4654.

Hammer, Ø. (2017) ‘PAST Paleontological Statistics, ver 3.17’, *University of Oslo, Oslo*, (1999), pp. 1–152. Available at: https://folk.uio.no/ohammer/past/past3manual.pdf.

Kleinbaum, D. G. and Klein, M. (2002) *Logistic Regression A Self-Learning Text Second Edition, Survival*. Available at: http://www.fao.org/tempref/AG/Reserved/PPLPF/ftpOUT/Gianluca/stats/Logistic Regression, A Self-Learning Text, 2Ed (Statistics For Biology And Health) (David G Kleinbaum, Mitchell Klein) 0387953973.pdf (Accessed: 2 August 2019).

Lehner, S. and Savage, K. (2008) ‘The effect of As, Co, and Ni impurities on pyrite oxidation kinetics: Batch and flow-through reactor experiments with synthetic pyrite’, *Geochimica et Cosmochimica Acta*. Elsevier Ltd, 72(7), pp. 1788–1800. doi: 10.1016/j.gca.2008.02.003.

Matheson, G. D. and Quigley, P. (2016) ‘Evaluating pyrite-induced swelling in Dublin mudrocks’, *Quarterly Journal of Engineering Geology and Hydrogeology*, 49(1), pp. 47–66. doi: 10.1144/qjegh2014-103.

Murray, J. and Henry, T. (2018) ‘Waulsortian Limestone: Geology and Hydrogeology’, *The International Association of Hydrogeologists Congress 2018*, (Session III), pp. 1–10.

El Naqa, I. and Murphy, M. J. (2015) ‘What Is Machine Learning?’, in *Machine Learning in Radiation Oncology*. Cham: Springer International Publishing, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

Newbury, D. E. and Ritchie, N. W. M. (2015) ‘Performing elemental microanalysis with high accuracy and high precision by scanning electron microscopy/silicon drift detector energy-dispersive X-ray spectrometry (SEM/SDD-EDS)’, *Journal of Materials Science*. Springer US, 50(2), pp. 493–518. doi: 10.1007/s10853-014-8685-2.

Onuk, P. *et al.* (2017) ‘Development of a Matrix-Matched Sphalerite Reference Material (MUL-ZnS-1) for Calibration of In Situ Trace Element Measurements by Laser Ablation-Inductively Coupled Plasma-Mass Spectrometry’, *Geostandards and Geoanalytical Research*, 41(2), pp. 263–272. doi: 10.1111/ggr.12154.

Paton, C. *et al.* (2011) ‘Iolite: Freeware for the visualisation and processing of mass spectrometric data’, *Journal of Analytical Atomic Spectrometry*. The Royal Society of Chemistry, 26(12), pp. 2508–2518. doi: 10.1039/c1ja10172b.

Pawlowsky-Glahn, V. and Buccianti, A. (2011) *Compositional Data Analysis*. Edited by A. Pawlowsky-Glahn, V. and Buccianti. Wiley. Available at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119976462.

Rodriguez-Galiano, V. F. *et al.* (2012) ‘Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture’, *Remote Sensing of*

*Environment*, 121, pp. 93–107. doi: 10.1016/j.rse.2011.12.003.

Sack, P. J., Large, R. R. and Gregory, D. D. (2018) 'Geochemistry of shale and sedimentary pyrite as a proxy for gold fertility in the Selwyn basin area, Yukon', *Mineralium Deposita*, 53(7), pp. 997–1018. doi: 10.1007/s00126-018-0793-5.

Sevastopulo, G. D. and Wyse Jackson, P. N. (2001) *Carboniferous (Dinantian) in The Geology of Ireland*. Edited by C. H. Holland. Edinburgh: Dunedin Academic Press. Available at: https://mail.google.com/mail/u/0/#inbox/FMfcgxwCgzDKgxtBgpKvZWhsjKlrSZlg (Accessed: 12 June 2019).

Strogen, P., Jones, G. L. and Somerville, I. D. (1990) 'Stratigraphy and sedimentology of lower carboniferous (Dinantian) boreholes from West Co. Meath, Ireland', *Geological Journal*, 25(2), pp. 103–137. doi: 10.1002/gj.3350250204.

Tuohy, B. (2012) 'Report of the Pyrite Panel', *Report of the Pyrite Panel*, June(June), p. 200. Available at: http://www.environ.ie/en/PyriteReport/FileDownLoad,30735,en.pdf.

Wang, L. J., Sawada, K. and Moriguchi, S. (2013) 'Landslide susceptibility analysis with logistic regression model based on FCM sampling strategy', *Computers and Geosciences*. Elsevier Ltd, 57, pp. 81–92. doi: 10.1016/j.cageo.2013.04.006.

Wilson, S. A., Ridley, W. I. and Koenig, A. E. (2002) 'Development of sulfide calibration standards for the laser ablation inductively-coupled plasma mass spectrometry technique', *Journal of Analytical Atomic Spectrometry*, 17(4), pp. 406–409. doi: 10.1039/b108787h.

Witten, I. *et al.* (2016) *Data Mining: Practical machine learning tools and techniques*. Available at: https://books.google.com/books?hl=en&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=Eibe+Frank,+Mark+A.+Hall,+and+Ian+H.+Witten+(2016).+The+WEKA+Workbench.+Online+Appendix+for+%22Data+Mining:+Practical+Machine+Learning+Tools+and+Techniques%22,+Morgan+Kaufmann,+Fourth+Edition,+2016&ots=8IFMwfozt8&sig=U7hT61ZkwhwST-4QvOb7oGMcStk (Accessed: 1 August 2019).

Xiong, Y. and Zuo, R. (2018) 'GIS-based rare events logistic regression for mineral prospectivity mapping', *Computers and Geosciences*. Elsevier Ltd, 111, pp. 18–25. doi: 10.1016/j.cageo.2017.10.005.

Yilmaz, I. (2009) 'Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat—Turkey)', *Computers & Geosciences*, 35(6), pp. 1125–1138. doi: 10.1016/j.cageo.2008.08.007.