



Improving Automated Geological Logging of Drill Holes by Incorporating Multiscale Spatial Methods

E. June Hill¹ · Mark A. Pearce¹ ·
Jessica M. Stromberg¹

Received: 8 July 2019 / Accepted: 27 February 2020 / Published online: 11 March 2020
© The Author(s) 2020

Abstract Manually interpreting multivariate drill hole data is very time-consuming, and different geologists will produce different results due to the subjective nature of geological interpretation. Automated or semi-automated interpretation of numerical drill hole data is required to reduce time and subjectivity of this process. However, results from machine learning algorithms applied to drill holes, without reference to spatial information, typically result in numerous small-scale units. These small-scale units result not only from the presence of very small rock units, which may be below the scale of interest, but also from misclassification. A novel method is proposed that uses the continuous wavelet transform to identify geological boundaries and uses wavelet coefficients to indicate boundary strength. The wavelet coefficient is a useful measure of boundary strength because it reflects both wavelength and amplitude of features in the signal. This means that boundary strength is an indicator of the apparent thickness of geological units and the amount of change occurring at each geological boundary. For multivariate data, boundaries from multiple variables are combined and multiscale domains are calculated using the combined boundary strengths. The method is demonstrated using multi-element geochemical data from mineral exploration drill holes. The method is fast, reduces misclassification, provides a choice of scales of interpretation and results in hierarchical classification for large scales where domains may contain more than one rock type.

✉ E. June Hill
june.hill@csiro.au

Mark A. Pearce
mark.pearce@csiro.au

Jessica M. Stromberg
jessica.stromberg@csiro.au

¹ CSIRO, 26 Dick Perry Ave, Kensington, WA 6151, Australia

Keywords Machine learning · Multivariate · Continuous wavelet transform · Edge detection · Tessellation

1 Introduction

Geologists are increasingly replacing or supplementing descriptions from visual drill core analysis (i.e. traditional geological logging) with numerical data from analytical devices to reduce inconsistency in information collection. However, to perform spatial prediction of geology (i.e. produce a space-filling three-dimensional geological model) from the information collected from mineral exploration drill holes, the geologist must impose a geological interpretation on numerical data. If the numerical data are manually interpreted, then the results will be subjective and depend on the experience of the geologist. Hence sets of results may be inconsistent when produced by different geologists or when collected and interpreted over a wide time interval and under various conditions. In addition, manual interpretation can be very time-consuming and challenging, especially when integrating multiple variables.

Automating or semi-automating the interpretation process using mathematical, statistical and machine learning (ML) algorithms is key to solving many of these problems. Using automation, the subjective input supplied by the geologist can be limited to the selection of suitable algorithms and parameters that can be recorded for future reference. The use of common algorithms and parameters for the entire data set ensures consistent results and allows experiments to be easily repeated. In addition, automation can provide rapid analysis of large and complex data sets.

In recent years ML has been recognised as a very powerful tool for dealing with high-dimensional geological data (i.e. large numbers of variables) and very large data sets (i.e. large numbers of samples). For example, Caté et al. (2017) used data from geophysical logs and compared a number of supervised ML algorithms to predict the probability of gold in drilling samples from a mine. Subsequently Caté et al. (2018) applied a number of supervised ML techniques to multi-element geochemical data in order to distinguish lithostratigraphic and alteration units in an ore deposit. It is interesting to note that the highest performing algorithms were not the same in the two different studies, reflecting the complexity and variability of the distribution of geological data in feature space. Fuzzy and probabilistic methods can also be applied to drill hole data. For example, Kitizig et al. (2017) demonstrated that the inclusion of petrophysical data with geochemistry (using unsupervised fuzzy *c*-means algorithm) can be useful for distinguishing rocks with similar chemistry but different textures and improve the overall classification rate. Silversides et al. (2015) used Gaussian processes (supervised ML) to provide probabilistic values to classify characteristic shale bands in iron ore.

ML has also been applied to geological data for maps and surface geology, for example, Cracknell and Reading (2014) and Ellefsen and Smith (2016). However, this is a very different spatial distribution of data and requires different spatial analysis to drill holes, which provide very dense data in one direction (i.e. down hole) and sparse data in all other dimensions. This is particularly the case for exploration drill holes which may be hundreds of metres or even kilometres apart.

ML algorithms developed for numerical data typically use a similarity measure based on distance between samples in feature space to categorise and classify samples. Drill hole data are spatial data, i.e. each measurement is taken at a specific location in space. However, this spatial information is not usually considered when applying ML to drill hole data. One reason is that it is unclear how to determine the best spatial scale to consider. Spatial geological information can be highly inconsistently applied when manually logging drill holes. This results in a range of approaches to logging from “splitters” (geologists whose logging is highly detailed) to “lumpers” (geologists who tend to group geological units). Results from ML algorithms applied to drill holes without reference to spatial information typically result in numerous small-scale units at the width of a single sample. These result not only from the presence of very small rock units, which may be below the scale of interest, but may also result from misclassification in two common forms (i) mixed samples and (ii) samples with a composition falling into the range of two or more rock types. Examples of these types of misclassifications will be shown in this paper. It has been demonstrated that incorporating spatial information into a ML algorithm can improve classification success (Hall and Hall 2017). However, it is not clear a priori what spatial scale to select, and this is critical to attaining usable results.

In previous studies, ML methods have been modified to include 3D spatial information for sets of dense drill hole data, such as brownfields exploration or mining situations. For example, the method of Fouedjio et al. (2017) uses geostatistical parameters to encode the joint spatial continuity structure of multiple variables, Romary et al. (2015) include spatial proximity as a condition for clustering and Bubnova et al. (2020) uses spatial data as a connectivity constraint for clustering. All these methods have been developed and tested for dense drilling situations and are less reliable in greenfields mineral exploration because of the large distances between drill holes. These methods require prior setting of spatial scale parameters (such as kernel size or similarity thresholds), and so provide a single scale result or a series of discrete scale results. Alternatively, non-ML methods can be used to analyse spatial information, the results of which may be used to domain drill holes into unclassified rock types, such as recurrence plots (Zaitouny et al. 2019, 2020). However, this method is not multiscale, as it also requires pre-setting of scale parameters linked to spatial scale.

In this paper a mathematical method is presented that allows rock type classification to be applied simultaneously across a range of spatial scales for each drill hole, see Fig. 1. This means that the geologist can select a suitable scale or set of scales for the problem at hand from the multiscale results. This new method, called Data Mosaic, groups samples of similar composition into spatially connected domains, which represent rock units. The domains are calculated for a range of spatial scales. The domains are hierarchical, in that larger scale domains are composed of progressively smaller scale domains. This scale hierarchy facilitates hierarchical classification, which allows a much richer classification than can be derived from the uniscale analysis provided by using machine learning algorithms on their own. The multiscale domains provide a framework to which any classification method can be applied, including many machine learning techniques, such as k -means, which is demonstrated in this paper. The domains are derived from geological boundaries detected using multiscale edge detection (continuous wavelet transform). Data Mosaic improves on the univariate

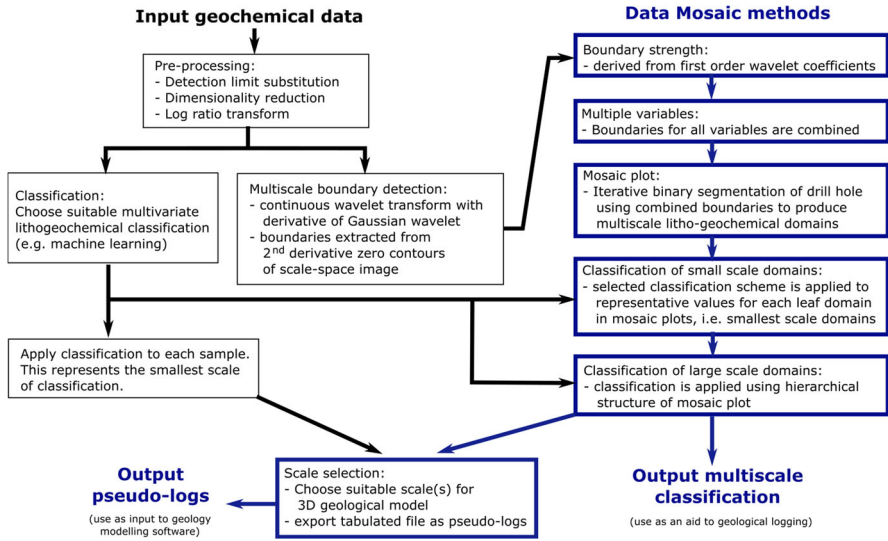


Fig. 1 Integration of multiscale analysis with litho-geochemical classification to produce pseudo-logs from geochemical data for 3D geological modelling. Text boxes with thin outlines indicate existing methods; text boxes with thick blue outlines indicate new methods introduced in this paper

multiscale wavelet tessellation method of Witkin (1983) and Hill et al. (2015) by providing a method for combining boundaries for multivariate data. It also uses a different measure of spatial scale to the original tessellation, which is demonstrated to be sensitive to both wavelength and amplitude of features in a regular way.

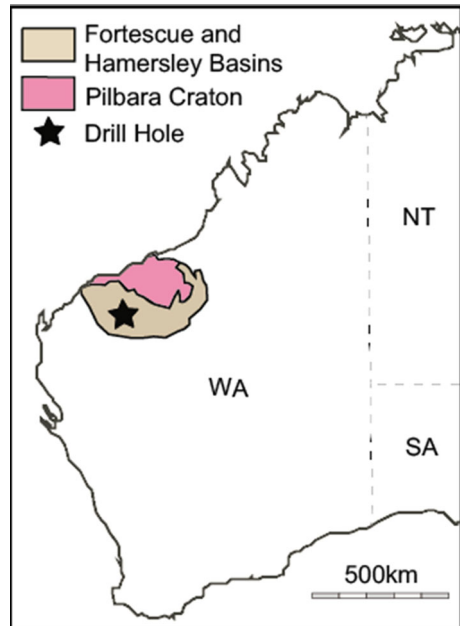
Data Mosaic is demonstrated using high spatial resolution X-ray fluorescence (XRF) data from a drill core that traverses sediments, volcanics and metamorphic rocks from the west Pilbara region of Western Australia. In addition, the ability to apply the method to large data sets is demonstrated using multi-element chemistry data from a densely drilled nickel-sulphide deposit in a layered mafic-ultramafic intrusion in Brazil. These data are described in detail in the Materials section. The Methods section describes the algorithms for multiscale boundary detection, estimating boundary depth and strength, combining boundaries to produce a multivariate, multiscale representation of the drill hole data (i.e. the mosaic plot) and, finally, applying a classification scheme to the mosaic plot (as outlined in Fig. 1). These methods are illustrated by application to the example drill hole data and to a simple synthetic data set.

2 Materials

2.1 Exploration Drill Hole, Western Australia

The drill core analysed in this study is from hole 18ABAD001 drilled by Artemis Resources through the stratigraphy of the Archean Fortescue Group of Western Australia (Fig. 2). The Fortescue Group is a sequence of sedimentary and volcanic rocks, up to 6.5 km thick. Drill hole 18ABAD001 is collared in volcanic and sedimentary

Fig. 2 Location of the case study drill hole in the Archean Fortescue basin of Western Australia



rocks of the Tumbiana Formation. The hole penetrates the Kylene Formation volcanics and Hardey Formation sediments before finishing in the felsic metamorphic rocks of the Pilbara basement.

The sediments and volcanics are part of a rift sequence, which began approximately 2775 to 2763 Ma and are described in detail by Thorne and Trendall (2001). The Hardey Formation sediments comprise up to 3 km of alluvial fan deposits passing into braided river and lacustrine rocks. The Kylene Formation volcanics erupted as a series of laterally extensive subaerial flows. Individual flows can be distinguished by the alteration of the porous and permeable flow top. The Tumbiana Formation sediments conformably overlie the Kylene formation; they comprise sandstones, argillites, tuffs and stromatolites (Thorne and Trendall 2001).

The drill hole was designed to delineate the geochemical stratigraphy and highlight areas of alteration that might be associated with fluid flow and mineralisation. Data collected included high precision laboratory bulk rock geochemistry and core-scale XRF scanning and spectral measurements. These complementary datasets were used to identify the key mineralogical and geochemical features of each rock unit and provide cross-references to ensure that the new core-scale XRF datasets provided accurate results. The XRF data were used in this study to demonstrate automated litho-geochemical core logging.

XRF scanning data was produced using a MinalyzeTM CS continuous XRF core scanner. The X-ray source was an Ag tube operated at 30 keV and 24 mA. Two certified reference materials (OREAS 24b and 624) were used to calibrate the beam flux and monitor instrument drift over the period of core scanning. An X-ray beam was moved along the core at a rate of 10 mm/s. The beam swath was 1 cm wide

and a fluorescence spectrum was recorded every second (every 10 mm). The XRF data were processed using XRS-FP software (Amptek Inc.). Spectra were binned at 1 m intervals to facilitate data visualisation and improve the signal to noise ratio.

The first stage of pre-processing involved imputation of missing values; these are left-censored missing values, i.e. below the detection limit of the analytical device. For this purpose, the R package *zCompositions* is used (Palarea-Albaladejo and Martín-Fernández 2015), which accounts for the compositional nature of the data. Prior to imputation all the data were processed to remove spurious correlations resulting from peak overlaps in the X-ray spectral deconvolution, removing all elements where more than 20% of the values were missing. *Minalyze*TM CS provides a detection limit value for each sample based on counting statistics of X-rays in the raw spectrum; the mean of the sample detection limits for each element was used as the detection limits for the imputation. A subset of elements was chosen for testing the automated geological logging methods described here, namely: Si, K, Cr, Rb, Sr, Zr and Ti. The subset elements were selected as they are known to be good discriminators of rock types in basaltic rocks (Cr, Zr, Ti) and can be used to identify variations in granitoid composition (Sr, Rb, K, Si).

The second stage of pre-processing applies a log-ratio transform to the compositional data prior to using any mathematical and ML methods. The additive log-ratio, ALR, is used (Aitchison 1986) by taking the log of each element ratio, using Ti as the denominator. This method is preferred to other log-ratio approaches because it is similar to the common practice of geochemists for normalising geochemical data with respect to Ti. This method produces variables that are familiar to geologists and that have geological meaning.

2.2 Large Data Set from Multiple Drill Holes

The Fazenda Mirabela intrusion is a mafic-ultramafic layered intrusion in Bahia, Brazil. A large set of multi-element ICPMS data was generated during the exploration of a nickel-sulphide orebody hosted by the intrusion. Analyses were performed on 1 m composite intervals from diamond drill core (Barnes et al. 2011). The data used in this example are from the west side of a boat-shaped intrusion where the layers dip to the east (Barnes et al. 2011). This example is used to illustrate the application of multivariate multiscale methods to large data sets. This subset contains 259 drill holes comprising almost 50,000 samples.

As in the Artemis data, the ALR transform was applied to selected elements from the data set using Ti (ppm) as the denominator. The selected elements (each concentration is measured in ppm) are: Al, Ca, Cr, Fe, K, Mg and Na. None of the samples in the data set were indicated as being below the detection limit; therefore, it is assumed that any substitution for missing data below the detection limit has already been performed.

3 Methods

3.1 Multiscale Boundary Detection

When geologists choose to manually segment drill core into rock units, they base their decision on the apparent thickness of the interval over which the unit exists, and on how distinctive the composition or physical properties appear. Therefore, scale selection depends on both wavelength (depth interval size) and amplitude (change in intensity) of the data signal. The method introduced here is based on boundary detection methods in image analysis and can incorporate information from both wavelength and amplitude of the drill hole data signal.

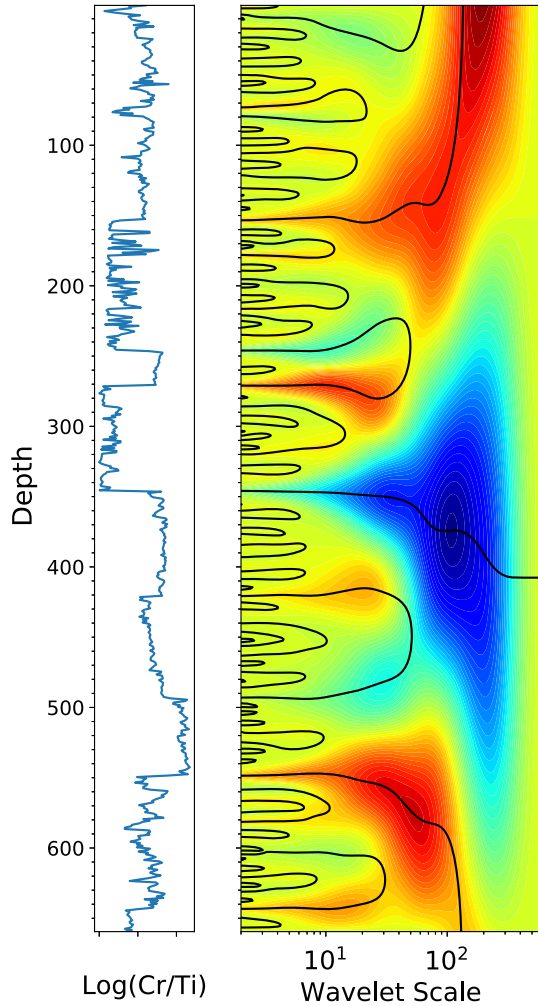
Boundary detection is a mature subject in image analysis, where it is more commonly termed edge detection (Marr and Hildreth 1980; Canny 1986). Mallat (1991) introduced the use of the continuous wavelet transform (CWT) as a fast and stable method for multiscale edge detection in images. In this method edges are extracted from the zero crossings of the second spatial derivative of the image using the derivative of Gaussian (DOG) wavelet (Mallat 1991; Mallat and Zhong 1992; Mallat and Hwang 1992; Mallat 2009). The zero crossings are the locations where the second derivative of the signal is zero; these locations represent inflection points in the signal. In image analysis the inflection points are generally considered to provide the best estimate of the edges in the image.

The CWT with DOG wavelet has been successfully utilised by geophysicists for detecting boundaries in well logs in the petroleum industry; for example, Cooper and Cowan (2009), Arabjamaloei et al. (2011), Davis and Christensen (2013) and Perez-Munoz et al. (2013). However, the scale-space plot (also known as a scalogram), which results from application of the CWT requires an expert to interpret it. Figure 3 shows an example of a scale-space plot of geochemical data (space is represented by depth down hole). In well logs the CWT method is primarily used to filter out small scale noise. The geophysicist produces a plot that illustrates the boundaries at a single scale determined by the operator as the best scale to represent the sedimentary facies; this is called blocking (Cooper and Cowan 2009). In blocking, all scale and hierarchical information is lost in the final product. The aim of the tessellation method (Hill et al. 2015) is to provide a multiscale product that is suitable for routine use by mining and exploration geologists and that preserves hierarchical information for classification, as described in the following sections.

The following pre-processing steps are performed prior to applying the CWT:

1. CWT requires data to be measured at constant intervals, data that do not comply are linearly interpolated (interpolation may result in some smoothing of extreme values).
2. In order to reduce signal edge effects, the signal is padded at the start and end of the data with a mirror image of the signal. The final signal length is a power of 2 for efficient use of the Fast Fourier Transform (FFT), with minimum padding of 1 signal length at either end. Different types of padding will give different results; the best padding to select will depend on an interpretation of the geology. It is

Fig. 3 Down hole values of log (Cr/Ti) and corresponding scale-space plot showing first order wavelet coefficients (colours: red = positive, blue = negative) and second order zero contour (black line)



important to recognise that coefficients of the CWT near the top and bottom of the drill hole signal may be affected.

The algorithm for applying the CWT with DOG wavelet used here is from Torrence and Compo (1998). The FFT of the DOG wavelet is calculated from Eq. (1) and normalised using Eq. (2)

$$\hat{\Psi}_0 = \frac{i^m}{\sqrt{\Gamma(m + 1/2)}} (s\omega)^m \exp(-(s\omega)^2/2), \tag{1}$$

$$\hat{\Psi}(s\omega_k) = \left(\frac{2\pi s}{\delta t}\right)^{1/2} \hat{\Psi}_0(s\omega_k), \tag{2}$$

where s is scale, m is the order of the derivative and δt is the sample interval length. The angular frequency, ω , is given by

$$\omega_k = \begin{cases} \frac{2\pi k}{N\delta t} & ; \quad k \leq \frac{N}{2} \\ -\frac{2\pi k}{N\delta t} & ; \quad k > \frac{N}{2} \end{cases}, \quad (3)$$

where N is number of samples. The normalised FFT of the DOG wavelet ($\widehat{\Psi}$) is convolved with the FFT of the signal (\hat{x}_k) using Eq. (4)

$$W_n(s) = \sum_{k=0}^{N-1} \hat{x}_k \widehat{\Psi} * (s\omega_k) \exp(i\omega_k n\delta t). \quad (4)$$

Finally, the inverse FFT is applied to the result. Wavelet scales are calculated as fractional powers of 2

$$s_j = s_0 2^{j\delta j}, \quad j = 0, 1, \dots, J, \quad (5)$$

$$J = \delta j^{-1} \log_2 \left(\frac{N\delta t}{s_0} \right). \quad (6)$$

Minimum scale (s_0) defaults to $2 * \delta t$ (Nyquist rate) unless the user wants to provide a factor larger than 2 to pre-filter the data. This can be useful for defining the smallest rock unit required in the classification and can significantly reduce computation time for high resolution data sets (e.g., wireline logging). The use of the FFT with CWT is mathematically efficient and computation time is very fast. Python's NumPy module provides methods for both forward and inverse FFT.

3.2 Locating Boundary Depth

The second order DOG wavelet is used in image analysis because it has two important properties (Mallat and Zhong 1992; Mallat 2009). First, it can detect inflection points in a signal (i.e. the zero crossings of the second derivative) used to estimate the location of edges in the image. Second, the zero crossings detected at larger scales will not disappear as scale decreases. This means that the zero contours in the scale-space plot can be traced back to the smallest scale. This allows the localisation assumption of Witkin (1983), which states that the true location of an edge in a signal at any scale is given by the location of the zero contour as the scale approaches zero. The mislocation of the edges at larger scales is caused by the smoothing of the signal during convolution with the stretched wavelet.

The edges detected by the CWT zero contours in numerical drill hole data are interpreted to represent geological boundaries. As shown by Hill et al. (2015), the scale-space plot can accurately locate geological boundaries by selecting the depth at which the zero contours intersect the smallest scale of the plot (Fig. 5). The precision of locating boundaries will depend on the depth interval of the measurements and the measurement noise in the signal.

3.3 Defining Boundary Strength

A drill hole can be segmented into different scale rock units (domains) based on the depths of boundaries and the boundary strength. The boundary strength is dependant on the amplitude and wavelength of the features. Hill et al. (2015) used the method of Witkin (1983) to determine boundary strength from the maximum scale of the zero contour on the scale-space plot (i.e. the black dot on Fig. 5). In this paper a new boundary strength measure is proposed, based on first order wavelet coefficients. This new method provides segmentation that is more intuitive as it directly accounts for amplitude of the features; in addition, it is more computationally efficient than the method of Hill et al. (2015).

The boundaries are extracted from the scale-space plot as follows:

1. Extraction of zero contours of the second order coefficients. Python's scikit-image measure module was used for this step.
2. The value of the first order coefficient that corresponds to each scale on the path of the second order zero contour is extracted.
3. Each zero contour that is not truncated by the edge of the plot is split into two parts at the maximum scale, to form a pair of zero paths. If a path has a first order coefficient of zero at the smallest scale, then it represents a horizontal inflection point this is not a boundary and is therefore omitted.
4. One boundary is defined from each zero path; each boundary has three attributes: (i) depth at the minimum wavelet scale; (ii) strength estimated from the maximum absolute first order wavelet coefficient on the path; (iii) polarity indicating whether the signal's gradient is positive or negative.

3.4 Calculating Multiscale Domains

Hill et al. (2015) used the tessellation method of Witkin (1983) to simplify the scale-space plot and provide a plot that can be interpreted by the geologist based on its similarity to a traditional geology log (Fig. 4). It differs from a traditional geology log in that, in addition to providing the location of geological boundaries, it also provides scale for the boundaries (i.e. maximum wavelet scale). Using the new boundary strength measure described here, a modified tessellation is calculated, which is faster to compute and does not include redundant domains resulting from horizontal inflection points.

Witkin's method results in a ternary tree structure; each zero contour forms a loop (except where truncated by the edges of the data) that divides the hole into 3 regions by depth: above, below and inside the contour (Fig. 5). With the new method, the ternary tree structure is replaced by a binary tree structure. This means that it is easier to deal with technical details, such as truncation of zero contours at the top and bottom of the hole, and ambiguities that occur when two zero contours intersect due to insufficient resolution in the scale-space plot.

The domains of the modified tessellation are calculated recursively, starting by splitting the root domain (i.e. a domain which spans the entire length of the drill hole signal) into two domains at the location of the strongest boundary. Each sub-domain

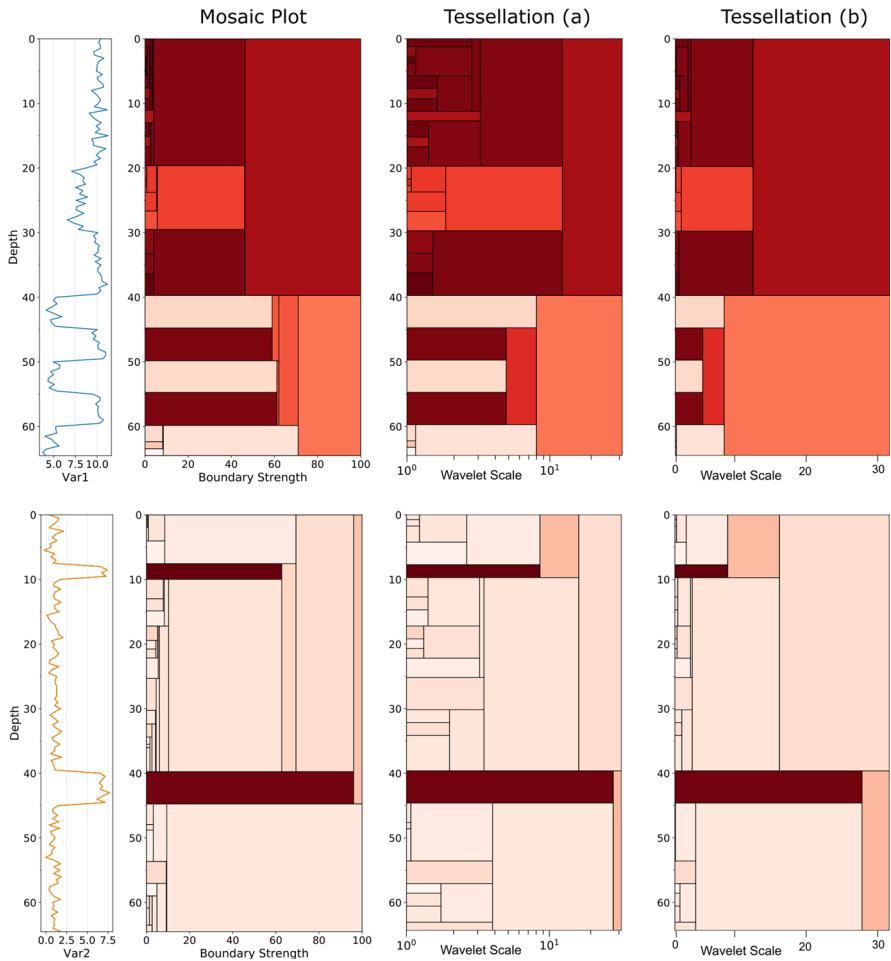


Fig. 4 Comparison between mosaic plot, first plot, (which uses wavelet coefficients as a boundary strength measure) and the tessellation plot, second and third plot, (which uses wavelet scale). The two tessellations are the same except one has a log scale for the x-axis (following convention) and the other has a linear scale for the x-axis (for easier comparison to the mosaic plot). The signals are synthetic rectangular signals with added Gaussian noise. The colours of the domains are proportional to the mean values of the signal over the depth range of each domain

is recursively split by the next strongest boundary within its depth range until all boundaries have contributed. The smallest scale domains are the leaf nodes of the binary tree structure. This method of producing a binary multiscale domain plot is called Data Mosaic. The method proceeds as follows:

1. Start with the largest (root) domain; i.e. the whole drill hole.
2. If the number of boundaries within a domain is greater than 1, split the domain into two using the strongest internal boundary and create two new rectangular domains, calculate the subset of boundaries that lie within each of the two new domains, and iterate for each new domain.

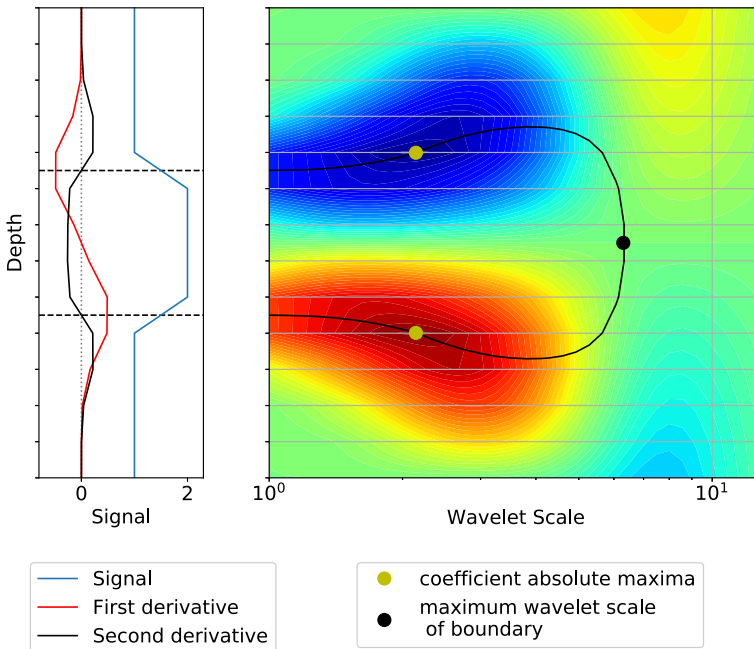


Fig. 5 Left plot: Simple synthetic signal (blue line) with 2 boundaries. The first order and second order derivatives for wavelet scale = 1 are shown as red and black lines, respectively. The dashed lines indicate the depth of the inflection points in the signal, i.e. the zero crossings of the second derivative. Right plot: scale-space plot of the CWT of the signal showing first order wavelet coefficients (colours: red = positive, blue = negative values) and second order zero contours (black curve)

3. If the number of boundaries in a domain = 1, add two domains as rectangles, with this boundary forming the splitting boundary; continue.
4. If there are no boundaries in the domain, this is a leaf domain; stop iteration.

At each splitting of the parent domain into two new rectangular sub-domains, the depth of the new domains comes from the top or bottom of the parent domain plus the depth of the splitting boundary. The maximum boundary strength of the domain is given by boundary strength of the splitting boundary; the minimum is temporarily set at zero. After all domains are calculated, the lower boundary strength of each domain is set by finding the next sub-domain in the list with the same depth for the top boundary of the rectangle and using the maximum boundary strength of the sub-domain. This simple method for finding the minimum boundary strength works because the binary tree structure is a depth-first traversal.

3.5 Combining Boundaries for Multivariate Data

Data Mosaic can be applied to any single numerical variable. However, for ML applications there is usually more than one variable to consider, therefore, the mosaic must be able to integrate multiple variables. To combine the wavelet transforms or even the

zero contours of multiple variables is an extremely challenging task, mainly because at higher scales the locations of the contours are distorted by neighbouring features. Combining the mosaic plots would result in a complex picture that may be very difficult to interpret. However, combining boundaries from each variable, using their depth and strength is a relatively simple task. A method is presented here whereby the boundaries are combined and a new multivariate mosaic is created from the combined boundaries. The steps for combining boundaries are as follows:

1. Boundaries are extracted for each variable for each drill hole.
2. Boundary strength is rescaled for each variable across all drill holes, for example, to [0, 1] or [0, 100], so the boundaries are comparable.
3. Boundaries for all variables are added into a single set of boundaries. If two boundaries are at the same depth or within a very small distance of each other (e.g., one interval length), then they are combined into a single new boundary. The depth of the new combined boundary is set at the depth of the strongest boundary. The strength of the new boundary (S_{new}) is increased according to the strength of all the contributing boundaries (S_i): $S_{\text{new}} = \sqrt{\sum S_i^2}$.
4. The polarities of all combined boundaries may be used to identify signal shape when determining a representative value for applying class labels, discussed in Sect. 3.7, below.

The CWT boundary detection method used here assumes that all boundaries are sharp. However, not all geological boundaries are sharp at the scale of sampling. For example, non-sharp boundaries may occur in drill holes where unaltered rocks are separated from intensely altered rocks by zones of increasing alteration or patchy alteration (for a more detailed discussion see Hill and Uvarova (2018)). For non-sharp boundaries the location of the boundary is at the inflection point of the smoothed curve traced back to the smallest scale. For non-sharp boundaries the inflection point will occur at low gradient changes, which are particularly susceptible to measurement noise in the data. This is a problem for multivariate data as the location of the boundary might be slightly dislocated for different variables. Once combined, this may result in multiple closely spaced boundaries instead of a single strong boundary. Pre-smoothing of the data may reduce this problem, but this also reduces the depth fidelity of the boundaries. It is possible that a more sophisticated method may be used to combine boundaries, especially for very high dimensional data, and is a potential subject for future study, perhaps using a probabilistic approach. The combination of boundaries assumes that the boundaries from each variable contribute equally. However, it is possible, although not yet tested, to weight the boundaries according to some criteria. Selection of appropriate weights may be a subjective method unless a supervised training approach is used.

3.6 Classification Using Machine Learning

Classification is independent of the Data Mosaic method, so any classification system can be applied including expert derived rules-based classification, unsupervised ML or supervised ML. The mosaic plot is used to provide the range of spatial scales for

the resultant rock units. Applying ML to the mosaic domains is demonstrated here using the simple and popular k -means clustering method. No particular clustering algorithm has been shown to best cluster rock types from drill hole data or to be the most useful for compositional geology data in general (Templ et al. 2008). This is not surprising, as the data structure will depend on the geological processes involved, which are many and complex. The k -means method was selected because it is often used as a benchmark ML algorithm for comparison to other algorithms and generally works reasonably well for large data sets.

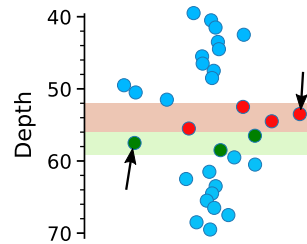
The k -means algorithm partitions the samples into k clusters, with each sample assigned to the cluster with the nearest mean. It is an iterative algorithm with a random starting state. The algorithm is generally considered to be fast but does not necessarily converge to a global optimum. A potential drawback of the method is that it expects clusters to be of similar size and approximately spherical in shape. For this example, clustering is performed with Python's `sklearn.cluster.k-Means` method with $k = 20$ clusters and default parameters. The number of clusters chosen is approximately equivalent to the number of rock types manually logged by the geologist. All variables are initially rescaled with zero mean and standard deviation of one (i.e. z -transformation).

3.7 Applying Classification to Multivariate Domains

Typically, the input to ML algorithms for numerical data is one vector per sample, where the length of the vector is equal to the number of variables, i.e. the dimension. To classify domains, a set of samples (i.e. vectors) must be considered and the set of samples can be of any length, depending on the depth extent of the domain. Hence it is not possible to directly import domain values into a clustering algorithm. If the set of samples in the domain forms a compact cluster in feature space, then the mean or median value could be considered as a reasonable representative value for the set. The median rather than mean value is preferred, as it will reduce the influence of mixed samples near the boundaries of the domain. Compact clusters may be a reasonable assumption for the smallest scale domains (i.e. the leaf nodes). However, the larger scale domains may contain data that groups more than one rock type, and therefore will likely be non-compact and possibly multimodal. So alternative methods that classify these domains must be explored. Hill and Barnes (2017) addressed this problem by using the symmetrical Kullback–Liebler divergence as a distance measure to compare sets of vectors. The Kullback–Liebler divergence (Kullback and Leibler 1951) can be used to compare the probability distributions of two sets of samples, irrespective of the shape of the probability distribution. However, the method can be computationally intensive for large data sets, and results depend on the resolution of the grid imposed on the probability distribution. Therefore, its practical application is limited to relatively small data sets.

An alternative approach to classifying larger scale domains is proposed, designed to mimic common practice of geological logging. Where geologists lump rock types, it is common for the geological log (i.e. the sequential record of the geology observed in drilling products) to contain two or more columns for rock type, so that multiple

Fig. 6 For narrow domains that represent maxima or minima in the signal (red and green regions, respectively), the extreme values (arrowed) are used as representative values instead of median values



rock types in one log interval can be listed in order of predominance. Similarly, for each domain, the rock types contained in its sub-domains can be recorded in order of predominance, where the most dominant rock type is the one that occupies the greatest total depth interval. The hierarchical structure of the multiscale domains in the mosaic plot provides all the information required to simulate this method of logging. The leaf nodes of the binary tree structure (i.e. smallest scale domains) that span the same depth interval as the domain under consideration contain the basic rock units of which the larger domain is composed.

One further type of domain needs to be addressed. This is a small-scale domain that forms a sharp maximum or minimum in the signal, where the median value may not be a good representation of the rock unit if a high proportion of the samples in the domain are mixed samples (Fig. 6). In such cases, the extreme value (i.e. the maximum or minimum) is considered to be the best estimate of the true composition of the narrow rock unit identified by the domain. These extreme domains satisfy the following criteria:

1. They are the lowest scale domains (i.e. leaf nodes of the binary tree structure).
2. The sample variability exceeds a specified threshold (default used is 85th percentile of the sample variability of all domains). Sample variability is a multivariate measure of dispersion which uses the mean Euclidean distance between the samples and the mean of the samples.
3. The domain size is less than a specified threshold (default used is five intervals).
4. The two domain boundaries are of opposite polarity (for at least one variable) indicating a maximum or minimum in at least one variable; see step 4 in Sect. 3.5.

3.8 Extracting Pseudo-Logs

The mosaic plot is designed to provide the geologist with the information needed for multiscale analysis in a format that is easy to decipher visually. The plot can be used as a geological logging aid by providing a visualisation of the multi-element data that is easy to digest. If the reason for the analysis is to construct a 3D litho-geochemical geology model, then the final step is for the geologist to select a suitable scale of rock type classification for generating their spatial models. The boundary strength is used as the measure of scale and can be applied across the data set to extract consistent scale pseudo-geological logs with hierarchical classification. The pseudo-logs can be

exported as interval data, similar in appearance to a traditional geological log, and imported directly into geological modelling software.

3.9 Programming

The method developed and presented here has been programmed using the Python language. Code is not available as it is commercially sensitive. However, access to the method described here is freely available for testing via a web application. Test data (including the data from Artemis Resources presented here) is provided as part of the web app, or the user may upload their own data. Please contact authors for access.

4 Results

4.1 Boundary Strength Measure

The intuitive nature of the new boundary strength measure is demonstrated using Figs. 7 and 8. These figures show the first order wavelet coefficients of two synthetic signals on scale-space plots. The plots illustrate how the maximum absolute values of the wavelet coefficients (and consequently boundary strength) increase as the amplitude and wavelength of the feature increases. By contrast, there is no observable regular relationship between the maximum wavelet scale of each boundary and the amplitude and wavelength of the feature.

4.2 Data Mosaic Plot

The mosaic plot is a modified form of the wavelet tessellation method. An example, using synthetic data, is used to illustrate the different results from the two methods; see Fig. 4. In most applications, wavelet scale is plotted using a log scale, therefore the tessellation is usually also plotted with a log scale on the x -axis. However a linear scale is probably more appropriate for comparing the tessellation with the mosaic plot, so both versions of the tessellation are illustrated. An example of the mosaic plot applied to a single variable from a real data set is illustrated in Fig. 9. The data mosaic method is almost twice as fast as the tessellation method (Hill 2017).

4.3 Multivariate Data Mosaic

Figure 10 demonstrates how boundaries are combined from multiple variables and the resultant multivariate mosaic plot. This example uses synthetic data in order to demonstrate the effectiveness of the method for retrieving the original domains. The figure also illustrates the effect of combining co-located boundaries to provide a larger overall boundary strength (compare boundaries on domain 6 with domain 8). The slightly stronger boundary between domains 9 and 10 compared to the apparently identical boundary between 7 and 8, is due to the edge effects from padding of the signal at the top and bottom of the hole. In the second example in Fig. 10, random

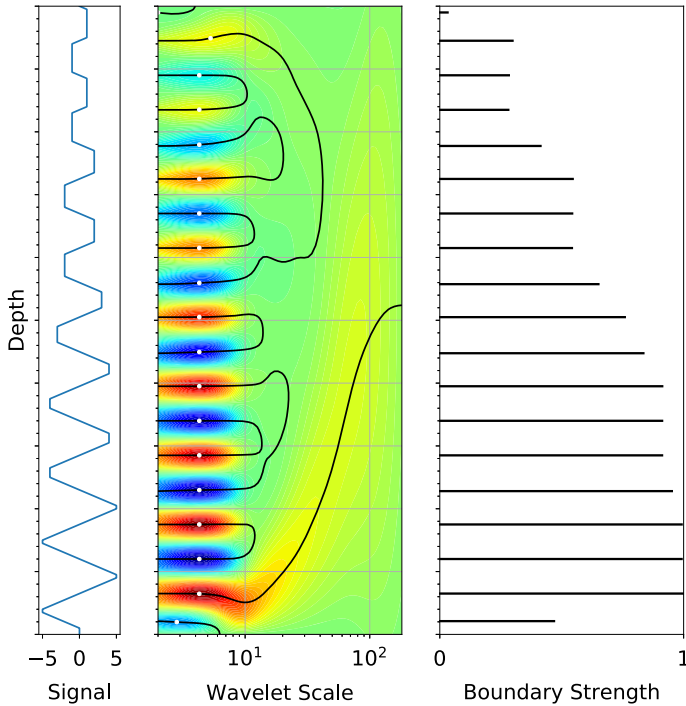


Fig. 7 Scale-space plot (middle figure) of synthetic signal (left) with varying amplitude; white dots indicate the locations of the absolute maxima of first order wavelet coefficients; right plot shows estimated depth and strength of boundaries

Gaussian noise is added to the signal to simulate rock heterogeneity or measurement error. In the lower mosaic plot in Fig. 10, it can be observed that the original domains can be retrieved at boundary strengths greater than 10. Small domains below this strength represent noise.

Figure 11 demonstrates detection of boundaries from multiple variables from real data and the combination of these boundaries. The results of applying *k*-means clustering to samples and domain representative values are illustrated in scatter plots, see Fig. 12. The number of clusters was chosen to be the same as the number of different rock types logged by the geologist, so that the results are comparable. The domains in the multivariate mosaic plot are coloured according to the classification (Figs. 13 and 14). In the case of leaf domains, these are the colours of the classes assigned using the *k*-means clustering. For non-leaf domains, the colour represents the dominant class in the set of leaf-domains that span the same depth interval (i.e. the lowest level children of the domain).

4.4 Classified Pseudo-Logs

Pseudo-logs have been extracted from the mosaic plot at scales of 0, 0.05, 0.1 and 0.25, shown as down-hole plots in Fig. 15. The information that can be exported as a

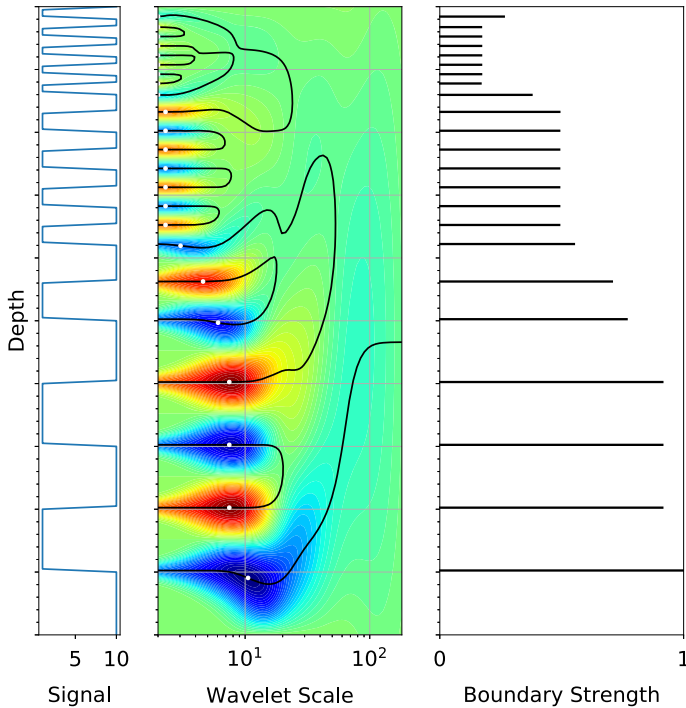


Fig. 8 Scale-space plot (middle figure) of synthetic signal (left) with varying wavelength; white dots indicate the locations of the absolute maxima of first order wavelet coefficients; right plot shows estimated depth and strength of boundaries (note edge effects in scale-space plot near start and end of signal)

pseudo-log in spreadsheet format is shown in Table 1. For this example, the number of exported rock types from the hierarchy has been restricted to three.

4.5 Large Data Sets

Automated methods are particularly valuable when dealing with large data sets due to their ability to produce a consistent interpretation in a very short amount of time. To demonstrate, 259 drill holes of multi-element geochemistry from the Mirabela intrusive complex, a total of 49,834 samples, were analysed. It took approximately 50 s on average to generate univariate multiscale domains for 7 variables for all drill holes, and a further 36 s on average to generate the multivariate domains using a Dell desktop computer with an Intel i7 processor. Parallelisation of the code would result in faster processing time, as each drill hole can be computed independently. Pseudo-logs have been extracted from the results for 2 scales, 0 and 0.1, and imported into geology modelling software for visualisation, Figs. 16 and 17.

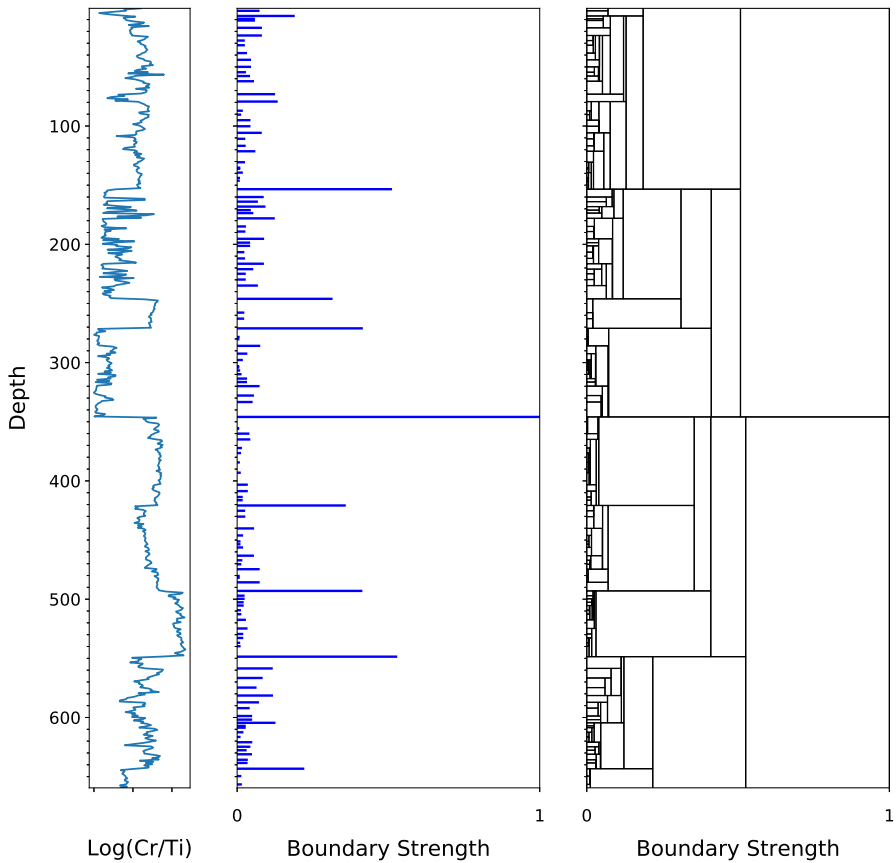


Fig. 9 Depth-corrected boundaries (middle plot) extracted from the $\text{log}(\text{Cr}/\text{Ti})$ values (left plot) with boundary strength plotted along the x -axis, and corresponding mosaic plot illustrating multiscale domains resulting from sequential binary partition of the space (right plot)

5 Discussion

5.1 Wavelet Coefficients for Multi-Scale Boundary Detection

This paper demonstrates the Data Mosaic workflow for interpreting multivariate numerical drill hole data as rock types, producing results that are suitable for generating three-dimensional geology models. The CWT is used to incorporate spatial information into the classification process. The CWT provides a fast method for distinguishing geological boundaries across a range of scales overcoming the need to know, a priori, the scale at which to select boundaries. In addition, the fast computation time of the CWT (using the FFT) makes the method suitable for processing large data sets, as demonstrated using the Mirabela data set of 259 drill holes. Using CWT, large numbers of drill holes can be analysed and interpreted in a consistent and objective manner.

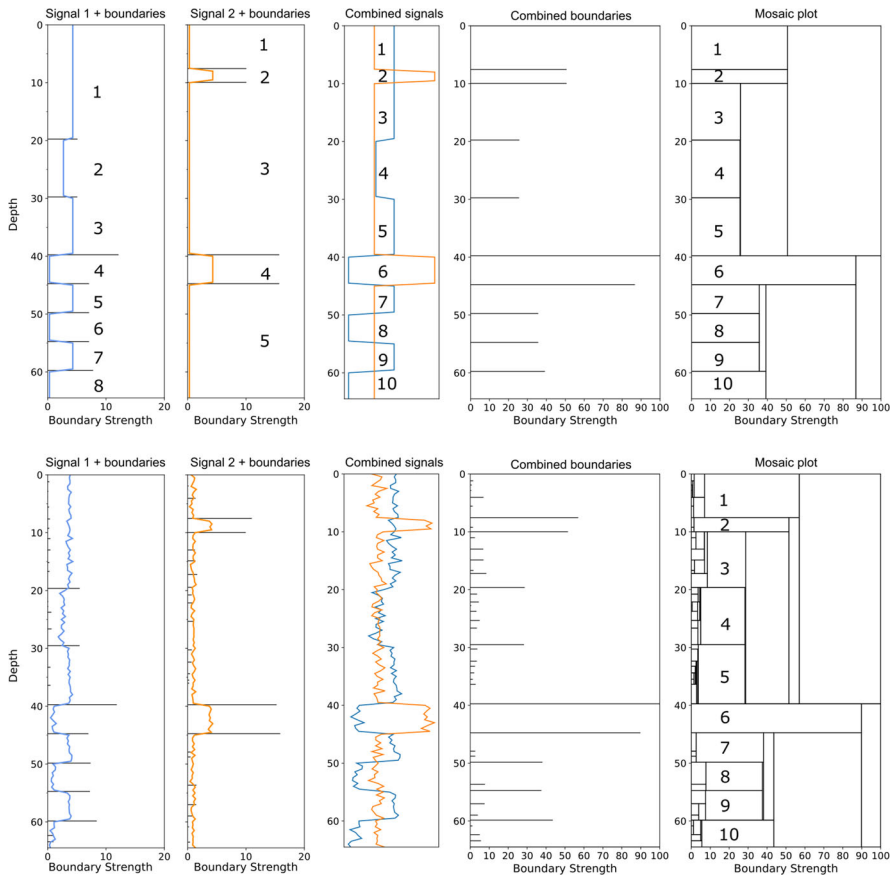


Fig. 10 Generating a mosaic plot using synthetic multivariate data. The top row of the figure shows boundaries and relative strengths for two simple rectangular signals (blue and orange). Boundary strengths are combined, then rescaled [0–100] and the mosaic is created. The bottom row shows the same method applied to the same signal but with added Gaussian random noise. Large numbers (1 to 10) indicate domains in the original and combined signals

Using a wavelet coefficient as a measure of boundary strength is useful for combining information from both amplitude and wavelength of features. Simple experiments with synthetic data show how increasing the wavelength or amplitude of features results in a corresponding increase in the absolute value of the wavelet coefficient (Figs. 7 and 8). Therefore, the wavelet coefficient provides a useful measure of geological scale by reflecting the apparent size of geological units as well as the degree of difference in composition of a geological unit to its neighbours.

The modified tessellation plot, which is based on wavelet coefficient instead of wavelet scale, results in a binary tree structure instead of a ternary tree structure. This simpler structure provides a multiscale domain plot, the mosaic plot, which is similar in appearance to the original tessellation (Fig. 4), but is faster to compute. However, the main advantage of the mosaic plot is, as described above, that the scale axis (i.e.

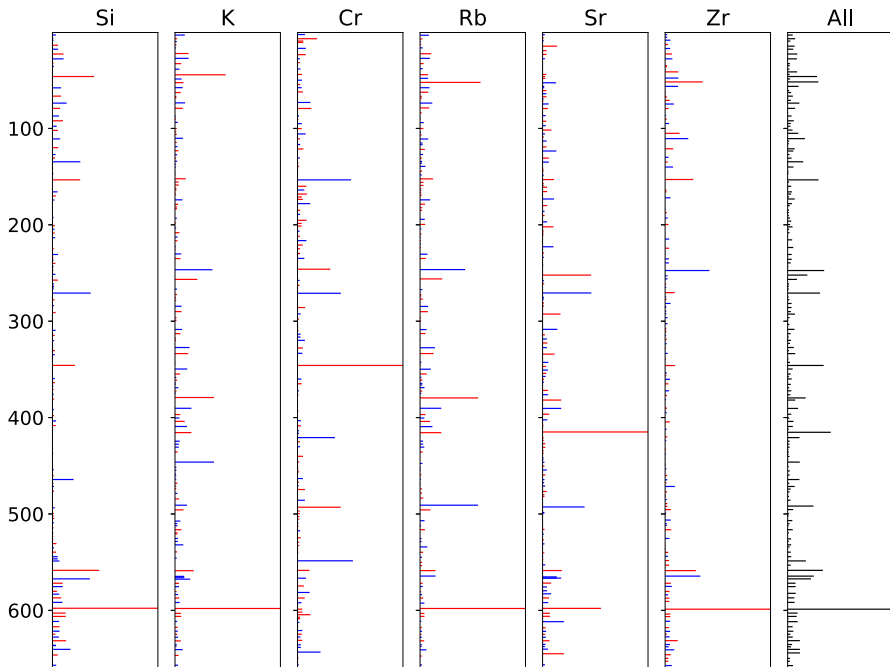


Fig. 11 Boundaries for a set of 6 compositional variables, plot labels indicate log ratios, e.g., Si = $\log(\text{Si/Ti})$, colours indicate polarity: red = increasing or blue = decreasing values; last plot shows all boundaries combined

boundary strength) has been demonstrated to be directly related to the amplitude and wavelength of features. The results shown in Fig. 4 also suggest that the noise is suppressed more effectively using the Data Mosaic method.

5.2 Multivariate Rock Type Classification Using Multi-Scale Spatial Information

Multiple variables can be integrated by combining their boundary locations and strengths, resulting in a multivariate mosaic plot. This method does not provide a multivariate classification for the data, instead it provides a multiscale framework for applying any classification system to the data. Therefore, it is not possible to compare the results to other classification algorithms. However, after applying a classification to the mosaic plot, the results can be compared to the same classification applied to the original samples to demonstrate upscaling and reduction of misclassification; see Fig. 14. The results can also be compared directly to the input data. Comparisons between the classified samples, the input data and the classified mosaic plot for example features is discussed in the rest of this section.

Visual comparison between the patterns in the data signals and the boundaries in the mosaic plot indicate how closely the boundary location and strengths are related to the signals. For example, in Fig. 14 the strongest boundary, close to 600 m depth,

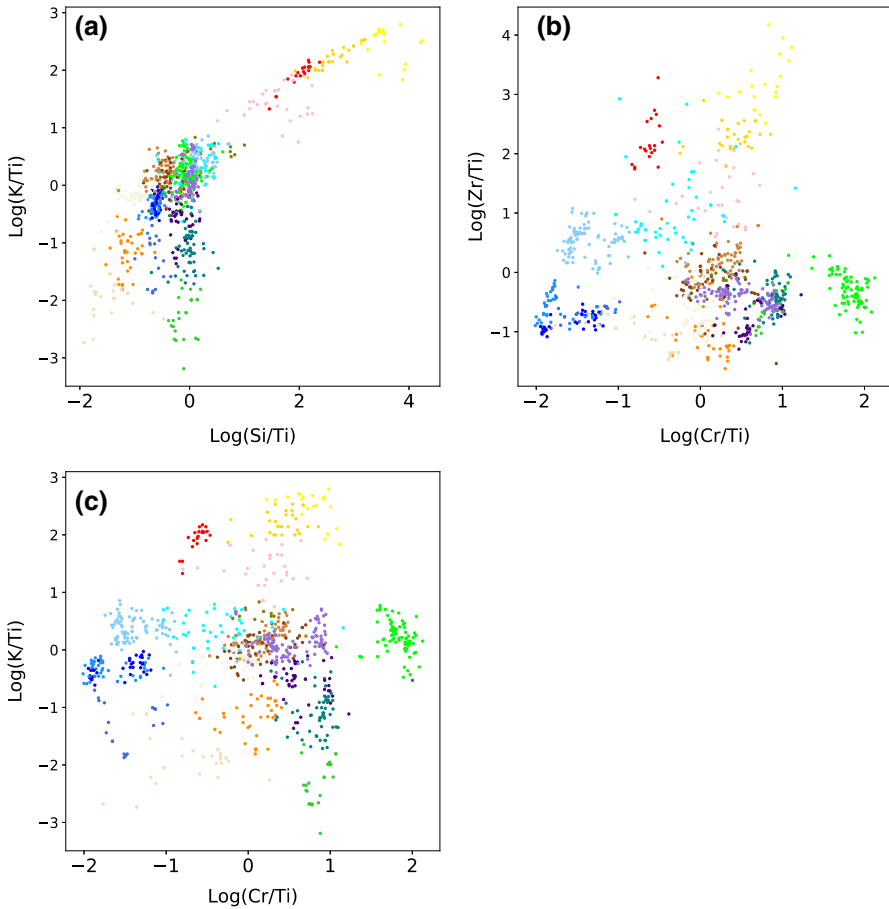
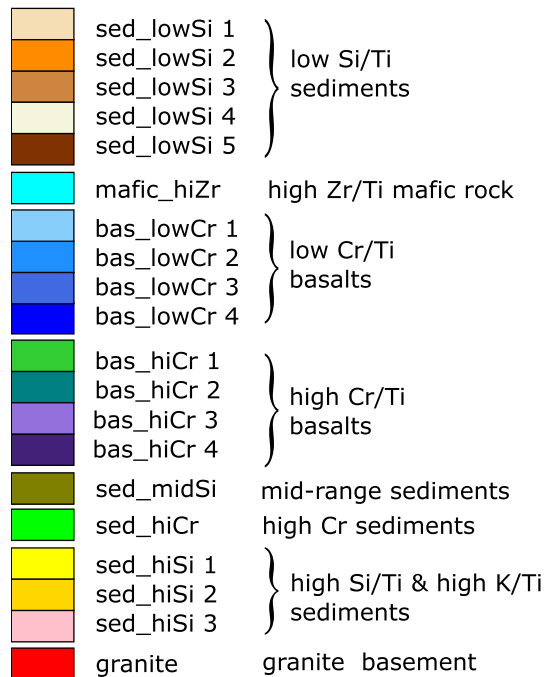


Fig. 12 Selected scatter plots showing the clustering results projected onto 2D space for three log ratio pairs, coloured by lithochemical classification

results from a change in rock type where all the variables except Cr/Ti show an abrupt change. This is the transition between low Si/Ti sediments and high Si/Ti sediments. By comparison, the boundary at approximately 490 m has a much weaker strength; although it forms a boundary between large wavelength features, a significant change in amplitude is visible in only two of the six variables; i.e. Sr/Ti and Cr/Ti. Very short wavelength variations such as the local spikes in composition (especially in Zr/Ti content) in the upper part of the hole result in fairly weak boundaries and narrow domains (narrow cyan coloured domains between 50 and 110 m, Fig. 14).

It is important to consider how the variables selected for analysis might affect the resulting combined boundary strength. For example, K and Rb concentrations are closely related, which will result in boundaries at similar depths with similar strengths (Fig. 11). By including both these variables in the analysis, the boundary strengths for boundaries in common are effectively doubled. This may be considered

Fig. 13 Legend for lithogeochemical classification of *k*-means classes; legend applies to Figs. 12, 14, 15



an undesirable result, in which case, only one of any set of closely correlated variables should be included in the analysis. The use of multiple, strongly correlated variables, however, would be useful for reducing the effects of noise, such as measurement error. Boundaries resulting from random noise will generally not coincide for different variables, and hence their strength, relative to common boundaries will be decreased.

A section of the drill hole is enlarged to illustrate the effect of upscaling the sample classification using the mosaic, Fig. 18. The smallest scale in the mosaic plot is approximately 2 samples, so isolated rock units which are a single sample interval in depth are merged with neighbouring samples. This is apparent in the individual cyan coloured units (i.e. mafic_hiZr) in the figure; where several of these units occur close together they are progressively merged into larger units in the mosaic plot. At even larger scales, the entire region from approx. 155 m to 245 m is merged into a single unit (i.e. bas_lowCr1), as this composite unit is distinctive from the composition of the large regions above and below it (see Fig. 12). The teal unit at approximately 245 m (i.e. bas_hiCr2) in the sample classification does not appear on the mosaic; it represents a mixed sample. At this location on the mosaic a single boundary marks the change from low Cr basalts to high Cr basalts, and the mixed rock sample is not included.

One aim of incorporating spatial information using the CWT is to reduce misclassification of mixed samples and samples that have uncertain cluster affiliation. Figure 19 shows several possible cases of misclassification caused by mixed samples. The signal plots (and the scatter plots Fig. 12) indicate that these highlighted samples

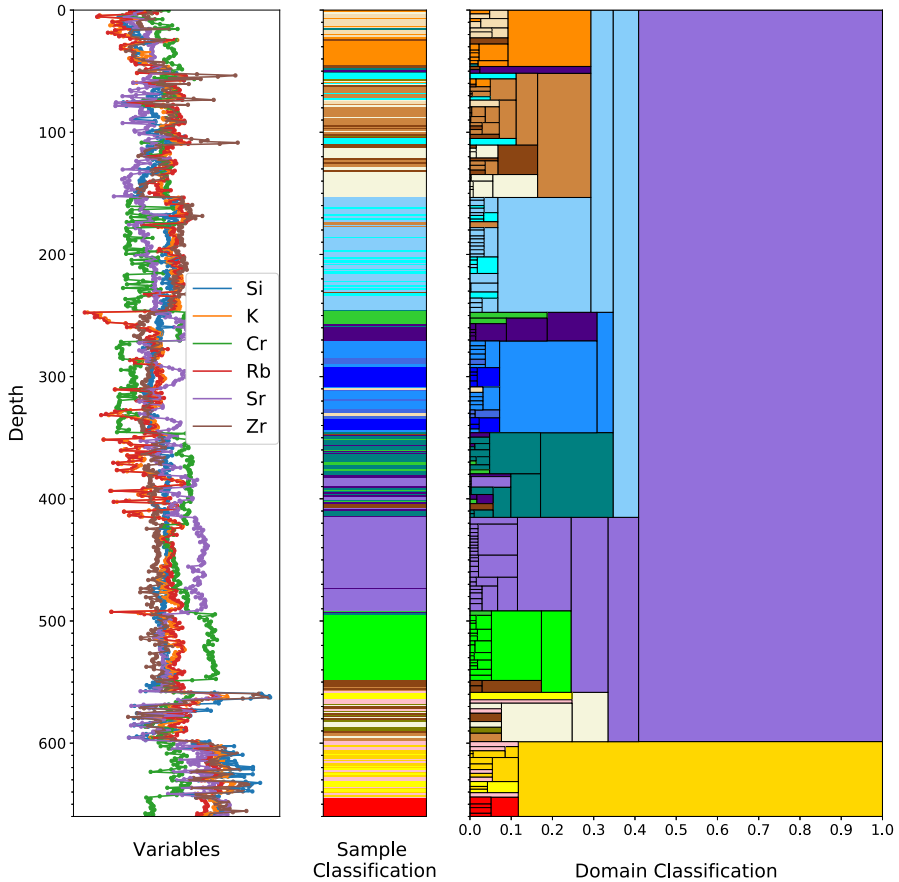


Fig. 14 Plots of scaled signals (left), samples coloured by litho-geochemical classification (middle), multiscale domains coloured by litho-geochemical classification (right)

lie compositionally between the two neighbouring units and are therefore likely to be mixed samples. These mixed samples are not shown as separate units on the multiscale domain plot because a single boundary will occur in the middle of the mixed sample, rather than on both edges of the sample. Figure 20 shows an example of a single sample which has been classified differently from all the samples in its neighbouring region, although the compositional difference between this sample and its neighbours is very small. This sample is unlikely to represent a distinct rock type. The possible misclassified unit is not represented on the multiscale domain plot because the compositional difference is too small to result in a significant boundary.

The domains identified in the mosaic plot can be compared with the geologist's log, Fig. 21. There is no expectation that the log and the mosaic will be identical because the input data are different, i.e. visual appearance of the core versus multi-element geochemistry. However, there is a strong correspondence between the major units in the geologist's log and the mosaic plot. For example, the amphibole porphyry

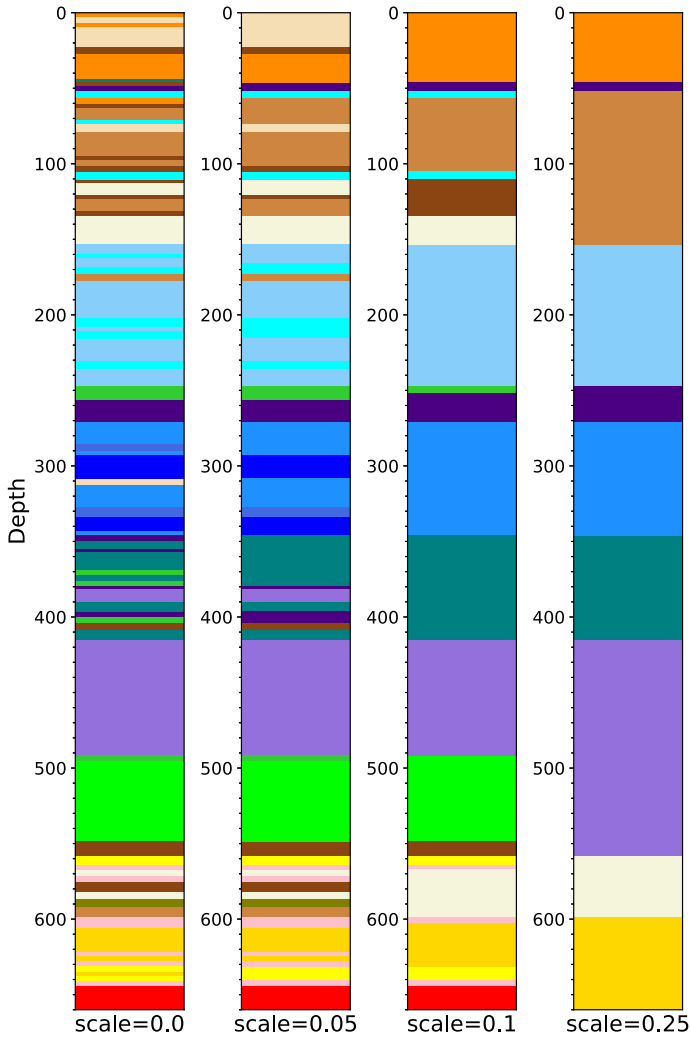


Fig. 15 Selection of pseudo-logs with increasing scale; the smallest scale from the domain plot (0) will reduce misclassification but may not include units which are less than 2 sample intervals in length; largest scale illustrated, 0.25 is probably the largest scale that is reasonable to consider before the grouping of rock types loses meaning

corresponds with the `bas_hiCr3` unit, the boundaries between the sedimentary package and the basalts are in the same location (approx. 152 m and 494 m). The geologist distinguished different types of basalt; however, from the description, these appear to be distinguished on the basis of textural differences (e.g. amygdaloidal, breccia) rather than geochemical differences, so the internal basalt boundaries do not correspond to the litho-geochemical boundaries in the mosaic. A similar observation can be made for the small sedimentary units, which are largely distinguished by grain size instead of chemistry. The boundary for the granite basement occurs in the same location in

Table 1 Format of exported file from pseudo-log at scale (boundary strength) of 0.1

Depth from	Depth to	Lith 1	Lith 2	Lith 3
0	46.16	sed-lowSi 2	sed-lowSi 1	sed-lowSi 5
46.16	51.77	bas-hiCr 4	sed-lowSi 5	
51.77	56.38	mafic-hiZr		
56.38	73.76	sed-lowSi 3	sed-lowSi 2	mafic-hiZr
73.76	105.09	sed-lowSi 3	sed-lowSi 5	sed-lowSi 1
105.09	110.59	mafic-hiZr		
110.59	134.6	sed-lowSi 5	sed-lowSi 4	sed-lowSi 3
134.6	153.38	sed-lowSi 4		
153.38	247.44	bas-lowCr 1	mafic-hiZr	sed-lowSi 3
247.44	252.16	bas-hiCr 1		
252.16	270.84	bas-hiCr 4	bas-hiCr 1	
270.84	345.91	bas-lowCr 2	bas-lowCr 4	bas-lowCr 3
345.91	379.59	bas-hiCr 2	bas-hiCr 1	bas-hiCr 4
379.59	415.25	bas-hiCr 2	bas-hiCr 3	bas-hiCr 4
415.25	491.8	bas-hiCr 3		
491.8	548.7	sed-hiCr	bas-hiCr 1	
548.7	558.48	sed-lowSi 5		
558.48	564.47	sed-hiSi 1		
564.47	567.31	sed-hiSi 3		
567.31	598.77	sed-lowSi 4	sed-lowSi 3	sed-lowSi 5
598.77	602.83	sed-hiSi 3		
602.83	631.41	sed-hiSi 2	sed-hiSi 3	
631.41	640.45	sed-hiSi 1	sed-hiSi 2	
640.45	644.15	sed-hiSi 3		
644.15	660	granite		

The predominant rock type for each domain is recorded in Lith 1; the following columns contain the other rock types that occur in the domain in reducing order of predominance

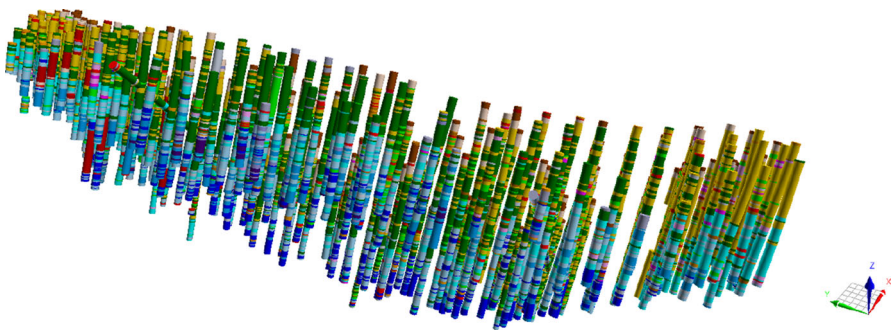


Fig. 16 Pseudo-logs of 259 drill holes shown at the smallest domain scale (0); 3D visualisation using Geoscience Analyst software; X, Y and Z are eastings, northings and depth, respectively

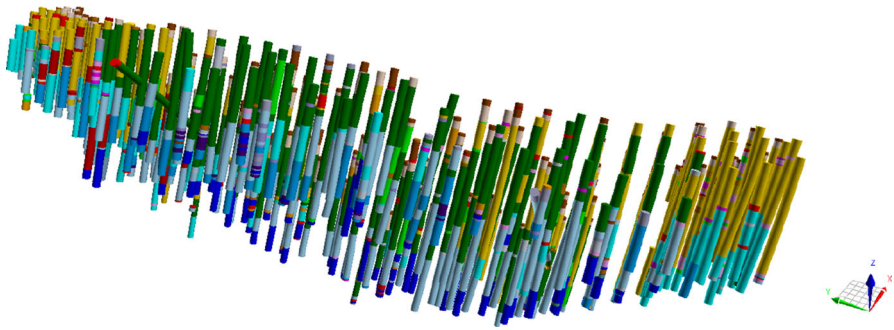


Fig. 17 Pseudo-logs of 259 drill holes shown at coarse domain scale (0.1), compare to Fig. 16; 3D visualisation using Geoscience Analyst software; X, Y and Z are eastings, northings and depth, respectively

both the geologist's log and the mosaic. In the mosaic this boundary is quite weak, indicating that the overlying sediments are similar in composition, perhaps suggesting that they are derived from the granite basement.

The mosaic plot of the geochemistry does not replace the geologist's log because they are different data sets. However, the mosaic plot provides added value to the geologist's log by providing the multi-element geochemistry data in a high-level visual format that is easy to digest. The geologist's log is highly detailed, on a similar scale to that of the sample classification. A second exploration hole has been drilled in this region and it is important to be able to correlate identical rock units between the two drill holes. This can be difficult when using very thin units, as thin geological units may not have sufficient lateral extent to span the distance between widely spaced exploration drill holes. In this case, the larger scale units in the mosaic plot could be useful for correlating similar composition units between drill holes, as these would be expected to have greater lateral extent. In addition, the non-subjective nature of the geochemical data can help correlate rock units that have been inconsistently labelled if they were logged by different geologists. Correlation of rock units is essential for building 3D geology models. The multiscale results also allow the geologist to select a suitable scale from the result to build a 3D geology model. For example, if the geologist required a regional model they might select a scale of 0.25, but for a more detailed model they might select a scale of 0.05, as illustrated in Fig. 15.

5.3 Multivariate Spatial Domaining For Hierarchical Rock Type Classification

The tree structure of the mosaic plot can be used to facilitate hierarchical classification of the multiscale rock type domains. The classification is initially applied to the leaf nodes of the binary tree (i.e. the smallest scale domains). Domains further up the tree structure inherit this classification information. This allows the mosaic plot to preserve information on mixed rock types in larger scale domains, which can be exported with the single scale pseudo-logs used for geology modelling. For example, hierarchical classification for a single scale pseudo-log is documented in Table 1. Using a hierarchical classification scheme overcomes issues with mixed rock types

Fig. 18 Expansion of Fig. 14 over approx. 160 to 260 m depth illustrating the effect of upscaling of sample classification by the mosaic plot. Mosaic plot is shown for boundary strength range of 0 to 0.3

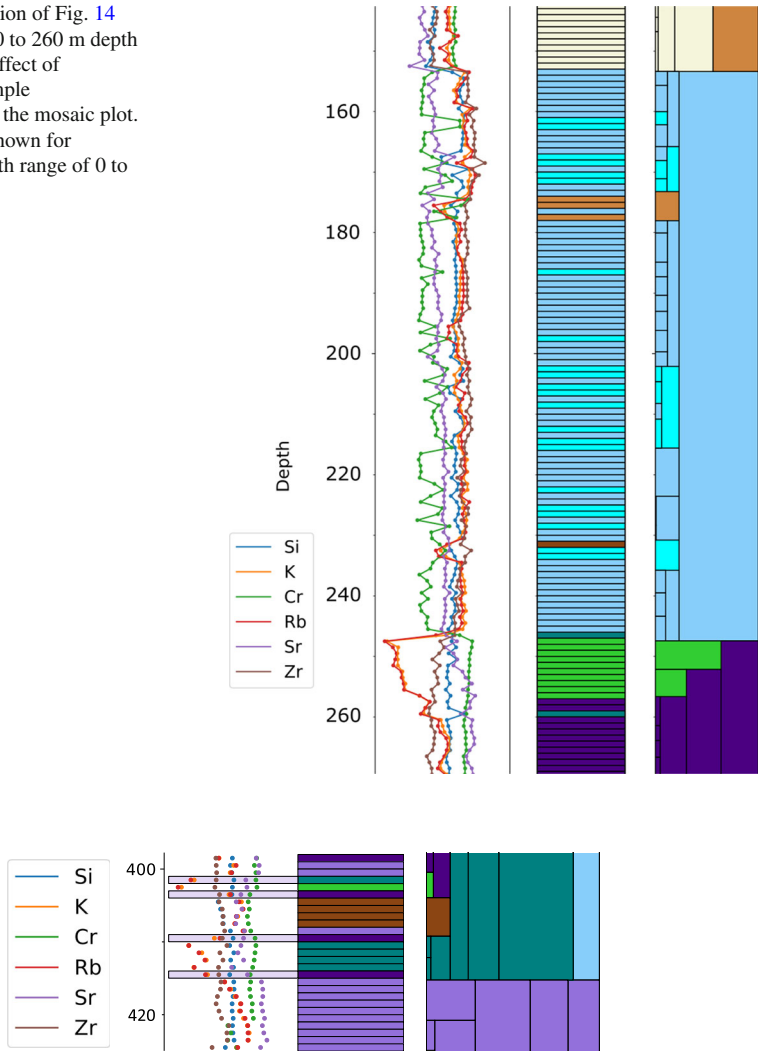


Fig. 19 Expansion of Fig. 14 over approx. 400 to 420 m depth showing probable mixed samples highlighted in light purple (left plot); sample classification (middle plot) and domain classification (right plot) are also shown for this depth interval

forming multimodal distributions of data that would be difficult to classify using expert rules-based classification or ML techniques.

The pseudo-logs provide a convenient method for importing hierarchically classified results into geology modelling software. They can be used to quickly test the effect of using different classification systems and different scales as inputs to a geology model. For example, larger scales might be preferred for stratigraphic or regional models, and smaller scales for orebody or deposit models.

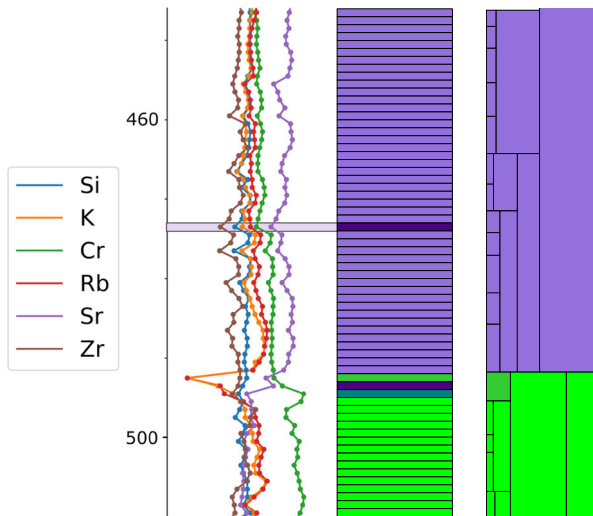


Fig. 20 Expansion of Fig. 14 over approx. 450 to 510 m depth showing probable misclassification (dark purple unit, middle plot) due to minor compositional change highlighted in light purple (left plot); domain classification (right plot) is also shown for this depth interval

5.4 Future Applications

The practical value of a geology model is enhanced if the level of model uncertainty is understood. In order to use stochastic simulation and information entropy to quantify uncertainty (Wellmann and Regenauer-Lieb 2012), it is essential to generate numerous scenarios that reflect the errors in the data and other sources of uncertainty. The ability to rapidly generate multiple solutions from the drill hole data, as demonstrated here, will make it possible to include uncertainty measures for geological boundary location which can be transferred to the model building software.

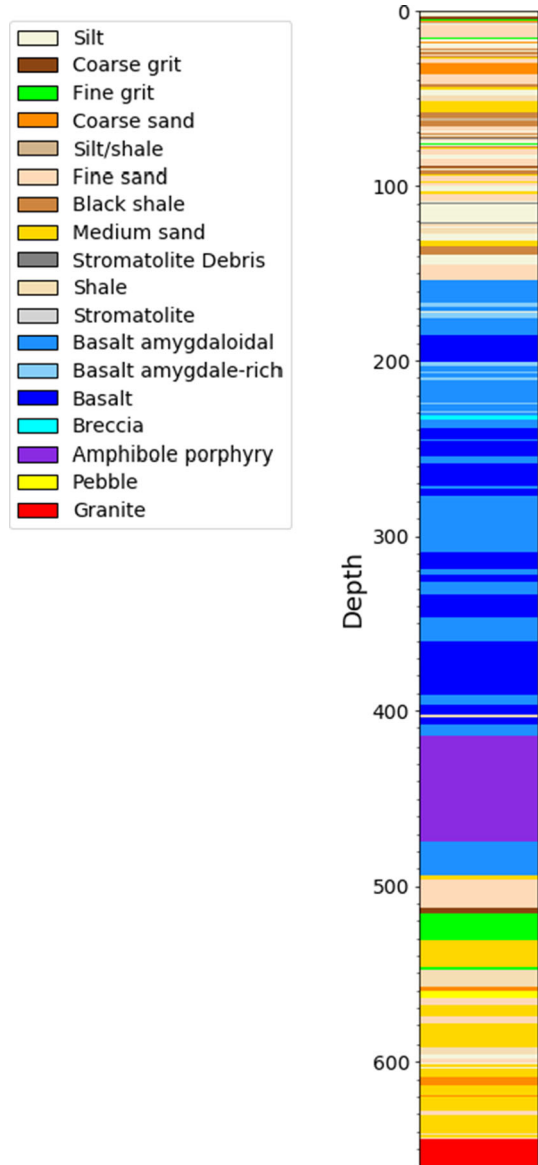
Automation and rapid processing has the additional advantage that it allows models to be updated with greater ease as new data become available. It also allows the user to test many different scenarios, including different combinations of variables, classification schemes and scales of evaluation. Furthermore, it facilitates a future where multiscale geology models can be generated.

5.5 Limitations of Current Application

The methods described here have been tested over a large range of data sets, mostly multi-element geochemistry data from mineral exploration drill holes, but also with wireline logging data from oil and gas wells and mineralogy from hyperspectral data. The method appears to have wide applicability, although it is essential to consider the input data limitations. These include:

1. *Sample density* The method is most valuable for high resolution data; it adds little value where the samples are sparsely distributed within the drill hole. The sampling

Fig. 21 Geologist's log of the drill hole



resolution needs to reflect the scale of variability that is required to be captured in the results and the precision of boundary detection needed.

2. *Sample interval* CWT assumes the data are sampled at regular intervals and requires that the data be interpolated and resampled if this is not the case. Therefore, it is not recommended for highly irregularly sampled data, as boundary detection depth precision will be variable.

3. *Sample spacing* For correct classification of rock units, at least one pure (non-mixed) sample is required. Therefore, the smallest rock intersection must be at least two sampling intervals in length. For example, to detect a rock unit of 1 m thickness, sampling must be less than or equal to 0.5 m intervals.
4. *Noisy data* Boundary depth precision depends on the smallest wavelet scale; i.e. s_0 in Eq. (5). However, noise from measurement error or natural compositional variation in rocks will affect the accuracy of the boundary depth estimate. For example, XRF measurement of light elements, such as Al, may be very noisy and therefore unreliable.
5. *Appropriate data types* Suitable data types are continuous numerical values of interval or ratio scale. The method is not suitable for ordinal data. The method assumes that all data values are meaningful. Therefore, the method should not be used to analyse data where missing values cannot be substituted by meaningful values. An example of this is the wavelength attribute in hyperspectral products, where the wavelength value is missing because the mineral is not present.

An important consideration is selecting variables for a suitable classification scheme. For data exploration tasks, the user may choose to include all variables. However, the results may be difficult to interpret. For more meaningful results we recommend selecting variables that represent geological processes: for example, using immobile elements for primary rock types and suitable element ratios which reflect alteration or weathering processes of interest. When using multivariate compositional data, it is important to use appropriate compositional data techniques for statistical analysis, such as the log-ratio transform.

Finally, the use of the DOG wavelet to detect boundaries implicitly assumes that the boundaries are sharp. If the boundaries are gradational in nature (i.e. gradational at the scale of sampling), then unexpected results may occur (Hill and Uvarova 2018). For example, gradational or patchy alteration may result in gradational boundaries in chemical elements or minerals at the scale of sampling. In all cases, the boundary will be located at the inflection point in the smoothed signal.

6 Conclusions

The CWT is a useful framework for detecting geological boundaries from drill hole data because it provides multiscale results that implicitly incorporate spatial information. The combination of local spatial information with compositional data results in the reduction of misclassification of rock types. In addition, the calculation of the CWT (using the FFT) is fast, so it is efficient for processing large data sets.

The wavelet coefficient is a useful measure for boundary strength because it reflects the apparent thickness of geological units and the degree of difference between a geological unit and its neighbours. Boundaries from multiple variables can be combined to form a multiscale, multivariate mosaic plot, where geological scale is specified by boundary strength. Classification can be applied to the mosaic plot in a hierarchical scheme.

The classified multiscale mosaic plot provides a visual tool for the geologist to select an appropriate scale of logging to generate a three-dimensional geological model that

is fit for its purpose. The pseudo-log extracted from the mosaic plot incorporates the hierarchical scheme of classification. The hierarchical classification overcomes difficulties in the automated classification of mixed rock types (e.g. multimodal distributions) that may occur when generating three-dimensional geological models at larger scales.

Acknowledgements Analysis of drill core 18ABAD001 was funded by the Australian Government Department of Industry, Innovation and Science through the Innovation Connections funding scheme. The authors would like to acknowledge Artemis Resources for allowing publication of results using their data. Thank you to Dr. Alex Otto, Dr. Andy Wilde, Dr. Andrew Rodger, Dr. Steve Barnes and Dr. Margaux Le Vaillant for testing the algorithm on many different data sets. In addition, we thank Dr. Jane Hodgkinson and other reviewers for many helpful comments and suggestions.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London
- Arabjamaloei R, Edalatkhah S, Jamshidi E, Nabaei M, Beidokhti M, Azad M (2011) Exact lithologic boundary detection based on wavelet transform analysis and real-time investigation of facies discontinuities using drilling data. *Pet Sci Technol* 29:569–578
- Barnes SJ, Osborne GA, Cook D, Barnes L, Maier WD, Godel B (2011) The Santa Rita nickel sulfide deposit in the Fazenda Mirabela intrusion, Bahia, Brazil: geology, sulfide geochemistry, and genesis. *Econ Geol* 106:1083–1110
- Bubnova A, Ors F, Rivoirard J, Cojan I, Romary T (2020) Automatic determination of sedimentary units from well data. *Math Geosci* 52:213–231
- Canny J (1986) A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 8:679–698
- Caté A, Perozzi L, Gloaguen E, Blouin M (2017) Machine learning as a tool for geologists. *Lead Edge* 36(3):64–68 (special section: data analytics and machine learning)
- Caté A, Schetselaar E, Mercier-Langevin P, Ross P (2018) Classification of lithostratigraphic and alteration units from drillhole lithochemical data using machine learning: a case study from the lalor volcanogenic massive sulphide deposit, Snow Lake, Manitoba, Canada. *J Geochem Explor* 188:216–228
- Cooper G, Cowan D (2009) Blocking geophysical borehole log data using the continuous wavelet transform. *Explor Geophys* 40:233–236
- Cracknell MJ, Reading AM (2014) Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput Geosci* 63:22–33
- Davis A, Christensen N (2013) Derivative analysis for layer selection of geophysical borehole logs. *Comput Geosci* 60:34–40
- Ellefsen KJ, Smith D (2016) Manual hierarchical clustering of regional geochemical data using a Bayesian finite mixture model. *Appl Geochem* 75:200–210
- Fouedjio F, Hill EJ, Laukamp C (2017) Geostatistical clustering as an aid for ore body domaining: case study at the rocklea dome channel iron ore deposit, Western Australia. *Appl Earth Sci Trans Inst Min Metall Sect B* 127(1):15–29
- Hall M, Hall B (2017) Distributed collaborative prediction: results of the machine learning contest. *Lead Edge* 36(3):267–269

- Hill J (2017) The data mosaic project: multi-scale spatial domaining of drill hole geochemistry data. Technical report EP171768, CSIRO
- Hill EJ, Barnes SJ (2017) Integrating spatial information and geochemistry for improved lithological classification of drill hole samples. In: Tschirhart V, Thomas MD (eds) Proceedings of exploration 17: sixth decennial international conference on mineral exploration, pp 853–856
- Hill EJ, Uvarova Y (2018) Identifying the nature of lithochemical boundaries in drill holes. *J Geochem Explor* 184:167–178
- Hill EJ, Robertson J, Uvarova Y (2015) Multiscale hierarchical domaining and compression of drill hole data. *Comput Geosci* 79:47–57
- Kitzig MC, Kopic A, Kieu DT (2017) Testing cluster analysis on combined petrophysical and geochemical data for rock mass classification. *Explor Geophys* 48(3):344–352
- Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Stat* 22(1):79–86
- Mallat S (1991) Zero-crossings of a wavelet transform. *IEEE Trans Inf Theory* 37(4):1019–1033
- Mallat S (2009) A wavelet tour of signal processing 3rd edition, chapter 6–wavelet zoom. Elsevier, Amsterdam, pp 205–261
- Mallat S, Hwang WL (1992) Singularity detection and processing with wavelets. *IEEE Trans Inf Theory* 38:617–643
- Mallat S, Zhong S (1992) Characterisation of signals from multi-scale edges. *IEEE Trans Pattern Anal Mach Intell* 14(7):710–732
- Marr D, Hildreth E (1980) Theory of edge detection. *Proc R Soc Lond* 207:187–217
- Palarea-Albaladejo J, Martín-Fernández JA (2015) zCompositions: R package for multivariate imputation of left-censored data under a compositional approach. *Chemom Intell Lab Syst* 143:85–96
- Perez-Munoz T, Velasco-Hernandez J, Hernandez-Martinez E (2013) Wavelet transform analysis for lithological characteristics identification in siliciclastic oil fields. *J Appl Geophys* 98:298–308
- Romary T, Ors F, Rivoirard J, Deraisme J (2015) Unsupervised classification of multivariate geostatistical data: two algorithms. *Comput Geosci* 85:96–103 (statistical learning in geoscience modelling: novel algorithms and challenging case studies)
- Silversides K, Melkumyan A, Wyman D, Hatherly P (2015) Automated recognition of stratigraphic marker shales from geophysical logs in iron ore deposits. *Comput Geosci* 77:118–125
- Templ M, Filzmoser P, Reimann C (2008) Cluster analysis applied to regional geochemical data: problems and possibilities. *Appl Geochem* 23:2198–2213
- Thorne AM, Trendall AF (2001) Geology of the Fortescue group, Pilbara Craton, Western Australia. Technical report 144, Geological Survey of Western Australia, Department of Minerals and Energy
- Torrence C, Compo GP (1998) A practical guide to wavelet analysis. *Bull Am Meteorol Soc* 79:61–78
- Wellmann JF, Regenauer-Lieb K (2012) Uncertainties have a meaning: information entropy as a quality measure for 3-D geological models. *Tectonophysics* 526–529:207–216
- Witkin AP (1983) Scale-space filtering. In: Proceedings 8th international joint conferences on artificial intelligence. Karlsruhe, Germany, pp 1019–1022
- Zaitouny A, Walker DM, Small M (2019) Quadrant scan for multi-scale transition detection. *Chaos* 29(10):103117
- Zaitouny A, Small M, Hill J, Emelyanova I, Clennell MB (2020) Fast automatic detection of geological boundaries from multivariate log data using recurrence. *Comput Geosci* 135:104362