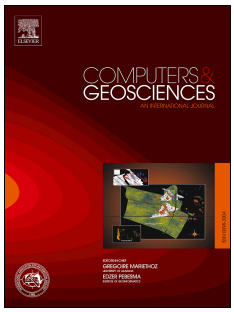
Accepted Manuscript



Geochemical characterisation of rock hydration processes using t-SNE

Tom Horrocks, Eun-Jung Holden, Daniel Wedge, Chris Wijns, Marco Fiorentini

PII: S0098-3004(18)30046-3

DOI: https://doi.org/10.1016/j.cageo.2018.12.005

Reference: CAGEO 4205

To appear in: *Computers and Geosciences*

Received Date: 17 January 2018

Revised Date: 31 August 2018

Accepted Date: 19 December 2018

Please cite this article as: Horrocks, T., Holden, E.-J., Wedge, D., Wijns, C., Fiorentini, M., Geochemical characterisation of rock hydration processes using t-SNE, *Computers and Geosciences* (2019), doi: https://doi.org/10.1016/j.cageo.2018.12.005.

Geochemical characterisation of rock hydration processes using t-SNE[1]

Tom Horrocks[1], Eun-Jung Holden[1], Daniel Wedge[1], Chris Wijns[1], Marco Fiorentini[1]
[1]Centre for Exploration Targeting, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009
Corresponding author: tom.horrocks@research.uwa.edu.au

ABSTRACT

Dimensionality reduction provides a simple, two-dimensional representation of multi-element geochemical assays, which facilitates visualisation of complex data and enhances their interpretation. A recently proposed dimensionality reduction algorithm, namely t-distributed stochastic neighbour embedding (t-SNE), generates effective two-dimensional representations of a wide range of datasets based on pairwise statistical distances of the input. However, direct application to multi-element geochemical assays has been shown to produce representations which can fail to separate specimens by a desired geological property, such as state of hydration. Since t-SNE is a statistical distance-based method, these sub-optimal representations may be due to the presence of dimensions (i.e., elements) irrelevant to the desired property—an issue often termed the 'curse of dimensionality'. To address this shortcoming, t-SNE was applied to (i) 31 elements in a geochemical assay database covering 16 000 drill core intervals intersecting the Kevitsa mafic-ultramafic intrusion (Lapland, Finland); and (ii) a subset of 11 elements capable of discriminating between unaltered and altered host rock specimens, as determined by a Random Forest classifier within a recursive feature elimination framework. The resulting representation more effectively separates altered and unaltered specimens, and we demonstrate that it produces more favourable representations than alternative well-known methods (namely, a self-organising map and principal components analysis) applied to the same dataset. We also demonstrate that the proposed t-SNE representation is applicable for re-logging of the specimens' alteration state as logged by geologists, and in particular provides visual insight into the labels suggested by a black box statistical re-logging algorithm.

[1] Mr. Tom Horrocks was responsible for experimental design, evaluation, and writing the manuscript. Prof. Eun-jung Holden and Dr. Daniel Wedge critically reviewed the manuscript with focus on computational elements. Dr. Chris Wijns provided details regarding the case study (Kevitsa). Both Dr. Chris Wijns and Dr. Marco Fiorentini critically reviewed manuscript with focus on geoechemical interpretation.

KEYWORDS

Dimensionality reduction, t-SNE, Geochemistry, Random Forest, Hydration, Feature selection

1 INTRODUCTION

Geochemical analysis, which decomposes geological specimens into their elemental concentrations, can refine the geological understanding of mineral deposits (Kyser et al., 2015). For example, rock units can be defined by clustering geochemical assays of drill core (e.g., Ellefsen et al., 2014; Meng et al., 2011), and predictive chemical models of lithology and alteration can be built with reference to corresponding geological interpretations (e.g., Cracknell et al., 2014). However, modern geochemical assays frequently contain concentrations for over fifty elements (Grunsky, 2010) and are thus considered high dimensional data, where the surplus of elements not only hinders effective visualisation, but also necessitates complicated statistical analysis. Dimensionality reduction is a solution to this problem, whereby the input data are transformed into a lower (often two) dimensional space, known as an *embedding*, which reveals the essential structure of the data (Hyvärinen et al., 2001). The embedding is visually interpretable and can be a noise-reduced basis for further analysis such as clustering (Templ et al., 2008; Reimann et al., 2008; Grunsky, 2010). Dimensionality reduction techniques recommended for geochemical assays are given by Grunsky (2010), which include: principal component analysis (PCA) (Hotelling, 1933; Pearson, 1901), multidimensional scaling (Torgerson, 1952), projection pursuit (Friedman and Tukey, 1974), independent component analysis (Hyvärinen et al., 2001), Sammon mapping (Sammon, 1969), and self-organising maps (SOM) (Kohonen, 1990).

Dimensionality reduction techniques can be divided into linear techniques such as PCA, which are computationally economical but not guaranteed to separate clusters (Chang, 1983); and nonlinear techniques such as Sammon mapping, which can theoretically represent nonlinear relationships between points but commonly produce lower quality embeddings on real-world data compared to PCA (van der Maaten et al., 2009). SOMs in particular have experienced increasing use in the geosciences, such as in order to characterise sedimentary provenance (Lacassie et al., 2004),

characterise phases of intrusive activity (Penn, 2005), integrate with other geoscientific data (Fraser et al., 2005; Fraser and Dickson, 2005), predict potential sources of airborne particulates (Gulson et al., 2007), and determine underlying processes contributing to water quality measurements (Juntunen et al., 2013). The SOM's embeddings are discretised (i.e., gridded) and the number of samples mapped to SOM grid cells may differ significantly which necessitates careful interpretation, although the grid structure can also be learned from the data with a growable cell structure SOM (Alahakoon et al., 2000) as demonstrated by Lacassie et al. (2004) and Lacassie and Ruiz-Del-Solar (2006).

A recently proposed nonlinear dimensionality reduction technique called t-distributed stochastic neighbour embedding (t-SNE) produces high quality and non-discretised embeddings which outperform many existing techniques on a variety of real-world datasets (van der Maaten and Hinton, 2008). Moreover, an approximated form of the algorithm extends its applicability to large ($n > 1000$) datasets (van der Maaten, 2014) such as deposit-scale geochemical studies. One such study demonstrated that t-SNE could separate mineralised and unmineralised specimens in an iron ore deposit, but could not produce adequate separation between hydrated and non-hydrated host rock specimens despite the presence of loss on ignition (Balamurali and Melkumyan, 2016), which is a strong indicator of hydration. In this study, we apply t-SNE to a large geochemical dataset (16 000 assays, 31 elements) of drill core intersecting the Kevitsa mafic-ultramafic intrusion in Finland which hosts a world class Ni-Cu-PGE deposit, with the primary objective of creating an embedding that clearly visualises the changes in elemental concentrations involved in host rock hydration.

This study extends the previous work by Balamurali and Melkumyan (2016) in three ways. First, we empirically identify a subset of elements that are jointly predictive of alteration, and use them to produce an embedding that separates hydrated and non-hydrated specimens. The degree of separation between hydrated and non-hydrated specimens is quantified based on the alteration status of each specimen's nearest neighbour in the embedding, and is shown to compare favourably to a t-SNE embedding generated using all elements, and an embedding based on PCA. Practical improvements of the t-SNE embedding over a SOM generated from the same data are also discussed. Second, we propose modifications to the t-SNE algorithm to address the compositional aspects of the input

geochemical data. Last, we demonstrate t-SNE's practicality for automated geochemical-based re-logging of alteration, where an embedding visualises the output of a black box re-logging algorithm as a function of the algorithm's user-defined parameters.

The remainder of this paper is structured as follows. In Section 2, the case study geochemical data and methods for element selection and dimensionality reduction are discussed. Section 3 presents the t-SNE embeddings generated from all elements and from the subset of selected elements, and the latter embedding is compared to those produced by PCA and SOM. A demonstration of how the t-SNE embedding can aid geochemical re-logging is also reported. Finally, conclusions are given in Section 4.

2 MATERIALS AND METHODS

Section 0 below provides details on the project area and the multi-element geochemical dataset used in this study. Section 2.2 describes the method for dimensionality reduction, which is summarised in Figure 1.

2.1 Case study

2.1.1 Geological setting

The Kevitsa Ni-Cu-PGE deposit—also known as the Keivitsa or Keivitsansarvi deposit—lies within a mafic-ultramafic intrusion hosted by the Savukoski Group of the Central Lapland Greenstone Belt, northern Finland. The intrusion is approximately 16 km^2 in surface area (Mutanen, 1997) and formed ca 2.06 Ga (Mutanen and Huhma, 2001). Mutanen (1997, pp. 135–139) separates the intrusion into three zones: a basal marginal chill zone (0-8 m), an ultramafic zone which hosts the deposit (up to 2 km thick), and a gabbro zone in the south-eastern part of the intrusion. Significant veining occurs throughout the deposit (Le Vaillant, 2014; Le Vaillant et al., 2016).

During regional greenschist facies metamorphism, the mafic minerals were hydrated into minerals including serpentine, amphibole, and talc (Mutanen, 1997). The olivine pyroxenite host rock underwent pervasive amphibole alteration, which was logged as metaperidotite by the mine-site

geologists (Gregory et al., 2011). The metaperidotite is generally accompanied by carbonate alteration, which is contained within the selvedge of nearby millimetre to metre-scale carbonate or carbonate-quartz veining (Gregory et al., 2011; Le Vaillant et al., 2016).

2.1.2 Data

The Kevitsa geochemical assay database (August 2014) in its unprocessed form contained 141 465 assays of exploration and grade control holes. A total of 51 elements were recorded in the database, however a mean of only 18 elements were present in each assay. A set of 31 regularly assayed elements were selected for further analysis: Ag, Al, Ars, Au, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, La, Li, Mg, Mn, Mo, Na, Ni, P, Pb, Pd, S, Sb, Sc, Sr, Th, Ti, V, Y, and Zn. The subset of assays containing all 31 previously listed elements described diamond drill core intervals (i.e., no grade control holes). These assays were then coupled with their corresponding geology logs, which contained optional fields for lithology (rock type), major and minor alteration, and type and degree of veining. The geology log depth intervals and assayed core depth intervals were not aligned, meaning that multiple geology logs often existed for one assay.

Assays were excluded from further analysis under the following circumstances. First, where the drill core interval was logged multiple times inconsistently (i.e., had overlapping geology logs with different lithologies). Second, where the drill core interval contained a vein, as the vein is volumetrically small and does not represent the intrusion alteration. Third, where the drill core interval was logged against a lithology not present within the intrusion, as the analysis was restricted to within the intrusion and not country rock. The final geochemical dataset comprised 16 165 chemical assays with a 31-element suite and no missing values. The final distribution of lithologies is described in Table 1, subdivided by whether the lithology is considered unaltered or altered.

2.2 Method

The dimensionality reduction method presented is summarised in Figure 1: first, replacement of rounded zeros in the geochemical assays (Section 2.2.1); second, an optional step of element selection

(Section 2.2.2); and third, multiple applications of t-SNE with only the lowest error embedding returned (Section 2.2.3). Each of these computational steps are discussed in turn below.

2.2.1 Rounded zero imputation

Many concentrations in the final subset of assays were marked as below detection limit (zero or negative of the detection limit). Unfortunately, substituting zero for these values excludes the application of logarithms, which is necessary for further analysis. A simple substitution to half the detection limit changes the covariance structure of the data (Martín-Fernández et al., 2011), which has an unpredictable effect in dimensionality reduction. To avoid these problems, the R package 'robCompositions' (Templ et al., 2011) was used to perform a model-based replacement of rounded zeroes (Martín-Fernández et al., 2012) using least squares regression and iterating until convergence. This method required predefined detection limits; in the absence of this metadata the largest negative number for each element was assumed to indicate the detection limit. In the case where no detection limits were indicated, the smallest measured positive number was used. These detection limits are given alongside the proportion of concentrations below detection limit in Table 2.

2.2.2 Element selection

The term 'element selection' refers to the process of empirically determining a subset of elements which discriminate between specimens according to an external geological property. To create embeddings which may better discriminate between hydrated and non-hydrated specimens, the specimens in the geochemical data were assigned labels of 'altered' or 'unaltered' according to their logged lithology (Table 1), and whether any alteration was explicitly logged against them. If a specimen had an unaltered lithology but had explicitly logged alteration, it was still considered altered. The labelled specimens were then used to determine a set of elements that were predictive of this alteration state using Random Forests within a recursive feature elimination framework, which are described in turn below. Note that a 'feature' refers to single dimension in the input data, which is an element in the context of geochemical datasets.

A Random Forest (Breiman, 2001) is a classifier composed of an ensemble of independently trained classification trees (Breiman, 1984) which aggregates the constituent trees' votes for a given input and classifies accordingly. The performance of the forest depends upon the strength of its trees, but also on their lack of vote correlation: the strength of the trees determines how frequently they cast the correct vote, while their low intercorrelation avoids a unilateral (possibly incorrect) vote. The method by which vote correlation is reduced is twofold: first, each tree is trained on a randomly resampled data set which covers approximately two-thirds of the original training set, created by sampling (with replacement) the original training set (i.e., bagging); second, the features used to split branches during training are chosen randomly.

Random Forests further leverage the bagged training set by evaluating each tree with its so-called *out-of-bag* samples, which constitute an unseen test set for that tree. The Random Forest calculates a 'feature importance' by randomly permuting values of a given feature between all out-of-bag samples on a tree-by-tree basis and calculating their average decrease in classification accuracy, known as 'out-of-bag accuracy'. Unfortunately, the feature importance is diluted between highly correlated features, as a highly correlated feature can compensate for the permuted feature with little resulting decrease in prediction accuracy. This effect can be mitigated by applying recursive feature elimination (Gregorutti et al., 2017), whereby the lowest ranking feature is iteratively removed with importances recalculated.

In this study, recursive feature elimination was applied to the labelled data. Feature importances were averaged from 20 Random Forests to reduce random effects. The Random Forests were trained with 1000 trees, and during bagging the unaltered specimens were subsampled to prevent class imbalance affecting the feature importances (10 704 unaltered specimens vs. 5461 altered specimens).

2.2.3 t-SNE

Given a set of assays from a (possibly reduced) set of elements, t-SNE was applied to produce two-dimensional embeddings.

2.2.3.1 Algorithm

Given a matrix of n D-dimensional points $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \mid \boldsymbol{x}_i \in \mathcal{R}^D)$, t-distributed stochastic neighbour embedding (t-SNE) aims to produce a corresponding matrix of low (typically two) dimensional points $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n \mid \boldsymbol{y}_i \in \mathcal{R}^K)$, $K < D$, where points that are similar in the original space are placed close together in the low dimensional space. For geochemical datasets, each point $\boldsymbol{x}_i$ represents a single assayed specimen with one dimension per elemental concentration. The joint probability p_{ij} defines the pairwise similarity (in the high dimensional space) between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and is defined as the mean of the pairwise conditional probabilities: $p_{ij} = \frac{1}{2}\left(p_{j|i} + p_{i|j}\right)$. While this definition of joint probability is unconventional, it has favourable characteristics over using conditional probability alone (see end of this section). The conditional pairwise probabilities are estimated using Gaussian kernels (Equation 1), where σ_i is the standard deviation of the Gaussian kernel for $\boldsymbol{x}_i$, and $\|\cdot\|$ denotes the ℓ^2-norm:

$$p_{j|i} = \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / \sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2 / \sigma_i^2\right)}. \quad (1)$$

Note that the raw similarity (the numerator) is normalised by all other raw pairwise similarities where $i \neq j$.

The standard deviation σ_i controls how quickly 'similarity' between two points decays as a function of their distance, and is dynamically computed such that similarity in regions of low density (i.e., where the closest neighbour is distant) decays more gradually. This is implemented by solving Equation 2 by binary search, where h is the user-defined parameter 'perplexity' that loosely corresponds to how many points should be considered highly similar to $\boldsymbol{x}_i$:

$$\sigma_i : \exp\left(-\sum_j p_{j|i} \ln p_{j|i}\right) = h. \quad (2)$$

In the two-dimensional embedding, the pairwise similarity q_{ij} between embedded points $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ is calculated using a Student t-distribution with one degree of freedom:

$$q_{ij} = \frac{\left(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2\right)^{-1}}{\sum_{k \neq l} (1 + \|\boldsymbol{y}_k - \boldsymbol{y}_l\|^2)^{-1}}. \quad (3)$$

The corresponding error C in the embedding is calculated as the Kullback-Leibler divergence between the joint probabilities p and q:

$$C = \sum_i \sum_j p_{ij} \log_2 \frac{p_{ij}}{q_{ij}}, \quad (4)$$

which corresponds to information lost (in bits) when q_{ij} is used to approximate p_{ij}. The algorithm randomly initialises the embedded points $\boldsymbol{y}_i$, and iteratively updates their positions by minimising C by gradient descent.

The improved performance of t-SNE over its forebear stochastic neighbour embedding is largely due to two factors. First, the new functional form for similarity between embedded points (Equation 3) still requires that similar points ($p_{ij} \gg 0$) are placed close together in the embedding ($\|\boldsymbol{y}_i - \boldsymbol{y}_j\| \approx 0$), but importantly does not constrain dissimilar points to be placed far apart. This enables embeddings which accurately model local structure in the high dimensional space. Second, defining similarity as $p_{ij} = \frac{1}{2}(p_{j|i} + p_{i|j})$ rather than just $p_{j|i}$ ensures that outlying points have a minimum similarity above zero: $\forall j \neq i, p_{j|i} \approx 0 \rightarrow p_{ij} \approx \frac{1}{2} p_{j|i}$. This penalises embeddings where outliers are proximal to other points, thus preferring outliers to be placed away from other points. These two improvements allow t-SNE to retain local structure while also isolating statistical outliers in the embedding.

2.2.3.2 Aitchison distance

The similarity functions that t-SNE uses rely on the Euclidean distance between points, expressed in terms of an ℓ^2-norm in Equation 1 and Equation 3 above. However, the Euclidean distance is a poor metric for comparing geochemical assays for two reasons. First, elemental concentrations are zero-bounded, typically log-normally distributed, and can have vastly different ranges; assays should at least be log-transformed or normalised lest the differences between major elements dominate those between minor and trace elements. Second, assays are compositional data (i.e., describe proportions of a whole), and therefore lie on a lower-dimensional simplex instead of occupying the full (half-)space.

Aitchison (1992, 1984) devised a distance function appropriate for data on a simplex based on elemental logratios, which can be written in two alternative but equivalent forms:

$$d_A^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{D}\sum_{i<j}^{D}\left(\log\frac{x_{1,i}}{x_{1,j}} - \log\frac{x_{2,i}}{x_{2,j}}\right)^2 \quad (5)$$

by Aitchison (1992), and

$$d_A^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{D}\left(\log\frac{x_{1,i}}{g(\boldsymbol{x}_1)} - \log\frac{x_{2,i}}{g(\boldsymbol{x}_2)}\right)^2 \quad (6)$$

by Aitchison (1986, p. 193), where d_A is the Aitchison distance, D is the number of dimensions (i.e., number of elements in the assay), $g(\cdot)$ is the geometric mean, and $x_{1,i}$ and $x_{2,i}$ are scalars from the i'th dimension of points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively. The difference between the Euclidean distance of the log-transformed data and the Aitchison distance can be obtained by rearranging Equation 6 to give

$$d_A(\boldsymbol{x}_1, \boldsymbol{x}_2)^2 = \|\log\boldsymbol{x}_1 - \log\boldsymbol{x}_2\|^2 - D\log^2\frac{\mathrm{g}(\boldsymbol{x}_1)}{\mathrm{g}(\boldsymbol{x}_2)}, \quad (7)$$

where $\log\boldsymbol{x} = (\log x_1, \log x_2, \cdots, \log x_D)$. The correcting factor $D\log^2\frac{\mathrm{g}(\boldsymbol{x}_1)}{\mathrm{g}(\boldsymbol{x}_2)}$ ensures that the Aitchison distance is scale invariant, that is, allows each point $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ to be scaled by different positive factors (e.g., $\alpha\boldsymbol{x}_1$, $\beta\boldsymbol{x}_2$) without changing value (Aitchison, 1992). This is necessary where the underlying data encodes size information about the specimen, either directly such as when measured in grams instead of units of density, or surreptitiously such as when mixing volumetric and mass density units (which may both be denoted in e.g. parts per billion). This correction is greatest between a point with roughly equal components and a point with many components close to zero; the Euclidean distance, Euclidean distance of log-transformed points, and the Aitchison distance on the simplex are compared geometrically in Figure 2(a-c), respectively.

2.2.3.3 Implementation

All embeddings presented in the following sections were computed alongside nine other embeddings with randomised initialisations and were selected for further analysis on the basis of embedding error (Equation 4). The Barnes-Hut t-SNE approximation (van der Maaten, 2014) as provided by the

'Rtsne' R package (Krijthe, 2015) was used due to the abundance of assays, as it requires only $O(n \log n)$ computation and $O(n)$ memory where n is the number of specimens. The trade-off parameter was set $\theta = 0.5$, which has been shown to work well on real-world datasets (van der Maaten, 2014).

All similarity calculations used the Aitchison distance in place of the Euclidean distance, which was implemented by applying an isometric logratio (ilr) transform (Egozcue et al., 2003) to the assays before supplying them to t-SNE: the Euclidean distance in the ilr-space is equivalent to the Aitchison distance in the original space. This enabled the use of existing t-SNE libraries without modification. Note that the transformed data occupies one fewer dimension even though the ilr transform is lossless; this is because the original data is restricted to a plane (i.e., a simplex) in the original space, but occupies the entire transformed space.

3 RESULTS AND DISCUSSION

This section presents a t-SNE embedding generated using all elements in the geochemical dataset and discusses the cluster structure present within the embedding. Following this, a t-SNE embedding is generated from the elements selected in Section 2.2.2 and compared to the previous embedding and alternative dimensionality reduction techniques in terms of separability of altered and unaltered specimens. Finally, a practical demonstration using an embedding to understand the output of a black box statistical re-logging algorithm is presented.

3.1 Embedding of all elements

Embeddings were first computed using all 31 elements and are presented in Figure 3, where each point represents one specimen (i.e., one drill core interval). The subfigures show the embedding coloured by the normalised log-concentration of each input element, using a 2% linear clip. The relevant ranges are given in Table 3 along with the true minimum and maximum for each element.

The embeddings produced by t-SNE can be used to visualise inter-cluster and intra-cluster structure with some caveats (Wattenberg et al., 2016): inter-cluster distances and the positions of the clusters in

the embedding are not necessarily informative, and the size of the clusters (in diameter) is not informative of the range of values within that cluster due to the dynamic adaptation of the similarity measure to sparse regions (Equation 2). Cognisant of these caveats, Figure 3 can be used to interrogate the qualitative degree to which each element controls the (distance-based) assay clustering.

Elements which cover a range of values across the cluster are unlikely to significantly contribute to the cluster structure. Such elements within a cluster appear either 'peppered' or are lain out in gradient bands across the cluster. Examples include gold, barium, cobalt, copper, iron, molybdenum, lead, palladium, sulphur, and silver to a lesser extent. These elements may benefit from further analysis, either independently or within a specifically-chosen subset of elements.

Elements which are highly bimodal (e.g., present at high concentrations or not at all) appear to strongly influence the division of clusters. For example, arsenic and thorium, which are either close to zero or are present in high concentrations (40% to 50% of values are below detection limit, see Table 2), are not clustered in mixed values. In some cases the clusters are not split (e.g., molybdenum), but the small-valued specimens are forced to one end of the cluster.

Figure 4 shows the embedding coloured by (a) the six most frequently logged unaltered lithologies ($n = 11\,940$), and (b) the three most frequently logged altered lithologies ($n = 3803$). Any specimens with lithology outside of the legend are shown in grey, which includes twelve minor unaltered lithologies ($n = 272$) and six minor altered lithologies ($n = 150$). The lithology colour scheme approximates olivine content (blue is low), and specimens which also have an entry under logged alteration (regardless of lithology) are plotted as triangles.

Overlaying the embeddings with logged lithology (Figure 4) shows that the cluster structure within the full suite of elements can only distinguish serpentinite (pink cluster in Figure 4(b)) and a combination of gabbro and uralite gabbro (blue clusters in Figure 4(a-b)). The embedding shows no discrimination between unaltered and altered lithologies. This does not indicate that there is no such discriminative information within the chosen set of elements, but rather that the elements jointly exhibit patterns that are unconnected to lithology – as previously discussed, the cluster structure

appears to be heavily influenced by the bimodal elements. An embedding with a subset of elements more suited to discriminating between altered and unaltered specimens is presented in Section 3.2.

Interestingly, almost all specimens which had some alteration logged against them (triangular points in Figure 4 were placed in several clusters towards the bottom of the embedding. Many elements contributed to this clustering: Figure 3 shows that vanadium, yttrium, titanium, scandium, strontium, calcium, aluminium, and cadmium are all present in uniquely elevated concentrations within this cluster. Silver and antinomy hold a tight range of intermediate values within this cluster, which is not seen elsewhere in the embedding. Still other elements are present in generally higher but more variable concentrations within this cluster, such as sodium and magnesium.

It should be noted here that the changes in element concentrations in Figure 2 are in fact dependent on changes of concentration for all the other elements. For example, an apparent elevation in scandium, which is not normally associated with alteration, can be explained by a reduction in many other elements that have been leached from the sample, such that a greater proportion of scandium remains, even though scandium itself has not been added by the alteration process. The tight grouping of specimens with logged alteration provides strong evidence for the validity of the alteration logging by geologists, although there are a few specimens with logged alteration positioned elsewhere in the embedding which may be strong candidates for re-logging.

3.2 Embedding of selected elements

Statistical distance-based analytic methods, such as k-means clustering or t-SNE embedding, can be influenced by which dimensions from the underlying data are used as input. This is because different clusters are apparent in the data depending on which dimensions are present, either due to true underlying structure or due to corruption by uninformative or noisy dimensions. This section reports how applying t-SNE to a subset of elements which were determined to be highly discriminative between altered and unaltered specimens improved the resulting embeddings in terms of separability between hydrated and non-hydrated specimens. The section concludes with a comparison with

existing dimensionality reduction techniques PCA and SOM, and an example of practical use for geochemical-based re-logging of alteration state.

3.2.1 Selected elements

Figure 5 shows the result of the recursive feature elimination used to select the subset of elements pertinent to alteration, where the y-axis shows the out-of-bag classification error of the Random Forests for discriminating altered and unaltered specimens using the number of elements in the x-axis. Each point in the figure is labelled by the current element with the lowest average feature importance, which is the next element to be eliminated. When read from left to right, the elements occur from least to most discriminative of altered and unaltered specimens. The classification error was judged to rapidly increase when fewer than eleven elements remained; hence, the eleven elements to the right of strontium (inclusive) were chosen for the alteration suite. In decreasing order of discriminative ability, these are: chromium, scandium, magnesium, yttrium, vanadium, manganese, calcium, aluminium, titanium, sodium, and strontium.

3.2.2 Embedding the selected elements

The embedding constructed from the eleven selected elements is given in Figure 6 (coloured identically to Figure 3). Figure 7 shows the embedding overlain with unaltered and altered logged lithology as previously in Figure 4, where a relatively strong separation can now be seen between the host rock olivine pyroxenite (yellow points in Figure 7(a)) and its amphibolised counterpart metaperidotite (green points in Figure 7(b)). In particular, the transition from olivine pyroxenite to metaperidotite can be seen to correlate with decreasing magnesium and manganese, and to a lesser degree with increasing calcium, vanadium, yttrium, and chromium (Figure 6, from the bottom-right to the top-left of the main cluster). The caveat explained in Section 3.1 regarding relative changes in element abundances applies here again: it is probable that yttrium was immobile during hydrothermal alteration, and the apparent increase in concentration is due to the removal of other elements.

This separation of hydrated host rock from non-hydrated host rock was completely absent from the previous embedding, perhaps due to the splitting of clusters by highly bimodal elements. This

separation was also absent in the t-SNE embedding presented by Balamurali and Melkumyan (2016), despite the presence of a strong indicator of hydration in the underlying assays (loss on ignition). This further demonstrates the importance of element selection prior to the application of t-SNE so that the salient elements (which in this case pertain to hydration) can sufficiently impact the resulting embedding.

Specimens with logged alteration (triangular points in Figure 7) are no longer completely disjoint from specimens without logged alteration as they were in the previous embedding (Figure 4). Manual inspection of the previous embedding shows elements such as cadmium and antimony clustered well in these points, but were not identified as discriminative by recursive feature elimination where they were eliminated first and fourth, respectively. This is because the order in which features are eliminated does not strongly indicate their individual discriminative ability, only that the strongest in a set of correlated features is removed last.

The improvement of separability between altered and unaltered specimens was quantified by relabelling each specimen according to its nearest neighbour in the embedding, and calculating the resulting accuracy (i.e., 1NN classification). It was found that the prior step of element selection improved 1NN classification accuracy on the embedding from 65.5% to 81.8% for altered specimens, and from 82.4% to 91.0% for unaltered specimens (Table 4). The improvement seen in the embedding is a direct consequence of achieving better separation in the higher dimensional space: element selection improved the 1NN classification accuracy on the original (ilr-transformed) data from 75.5% to 82.3% for altered specimens, and from 87.7% to 92.2% for unaltered specimens (Table 4). Note that better separability is to be expected in the full dimensional space (which cannot be directly visualised), since a low dimensional embedding can only approximate the inter-point distances in the original space. Indeed, the proportion of points which maintain their nearest neighbour is a measure of embedding quality (Sanguinetti, 2008; van der Maaten et al., 2009).

3.2.3 Comparison with other techniques

The separability of altered and unaltered specimens was compared with two existing techniques: visualisation of the first two principal components (Figure 8) and an SOM embedding (Figure 9).

Both techniques used the ilr-transformed assays of the 11 selected elements for direct comparison. For visual clarity, both figures are coloured by the alteration status used during element selection (i.e., 'altered' or 'unaltered') rather than the logged lithologies. It is important to note that the t-SNE embeddings derived from the selected subset of elements are being compared with embeddings from alternative techniques PCA and SOM on the same subset of elements. PCA and SOM are included here for comparison only; they are not being suggested for dimensionality reduction prior to application of t-SNE, nor are they related to the feature selection performed in this study.

Although PCA (Hotelling, 1933; Pearson, 1901) is a linear technique which is desirable for various applications, it is often used for dimensionality reduction due to its computational simplicity, freedom of user-supplied parameters, and interpretable output. We show here that it can be inadequate for this application where the data contain subtle, nonlinear chemical patterns: Figure 8 displays the first two principal components, where significant mixing between the altered and unaltered specimens can be seen. The nearest neighbour classification accuracy is 76.6% for altered specimens and 90.3% for unaltered specimens, which are both lower than for the corresponding t-SNE embedding (Table 4). Further scatter plots were produced to visualise all remaining pairs of principal components; some pairs were able to produce a small cluster of unaltered specimens, however, the remaining specimens (i.e., the majority) remained mixed as in Figure 8.

The SOM embedding (Figure 9) shares the important large-scale characteristics of the corresponding t-SNE embedding (Figure 6), namely a separation between altered and unaltered specimens characterised primarily by low magnesium and manganese concentrations (lower third of the right side of the SOM embedding), and a region corresponding to specimens with logged alteration (bottom-left corner of the SOM embedding). However, the discrete grid embedding that characterises the SOM technique hides individual outliers (peppered points in Figure 9(a)), and is more difficult to plot with multiple labels (each grid must show proportions of labels). Moreover, plotting the elemental concentrations on the SOM gives a false sense of scale: the high-manganese region covers approximately one third of the SOM cells but only accounts for a small number of points (Figure 6,

clusters to the top and bottom). Correct interpretation of the SOM requires reference to a 'hit chart' detailing the number of points assigned to each cell.

3.3 Example scenario: geochemical re-logging

The lithology and alteration of drill core at Kevitsa was logged by multiple geologists over a span of time and is subject to human error. Geochemical core re-logging is a quantitative method for correcting erroneous geological logging, whereby a classifier is trained and executed for prediction of geological logging from geochemistry. The classifier can be made to favour a simpler geochemical model over a more complex model with higher predictive accuracy on the training set by *regularisation*. The degree of regularisation is usually determined by maximising predictive accuracy on an unseen test set (rather than the training set); however, if the labels are incorrect then maximising accuracy leads to an incorrect model. In this section, a Random Forest is used to re-log the simplified alteration status (i.e., 'altered' or 'unaltered') and demonstrate how visualising the forest's classification on the embedding of selected elements from the previous section can be useful in determining an appropriate degree of regularisation.

A Random Forest with 1000 trees was trained to predict alteration status ('altered' or 'unaltered') from the geochemical data (all 31 elements); it was deemed unnecessary to use the feature-selected set of elements as the Random Forest performs internal feature selection during training. Although Random Forests are not prone to overfitting, they can be forcibly regularised by limiting the size of the constituent trees. By default, the trees within a Random Forest are grown to their maximum depth such that each leaf node describes one instance of the bagged training set. Enforcing a minimum terminal node size limits the size of the individual trees and thus regularises the forest. In this experiment, the minimum terminal node size of the Random Forest was varied between 1 (fully grown trees) to 8192 in a doubling sequence and the resulting predictions for all specimens were recorded and visualised on the 11-element embedding (Section 3.2).

Figure 10(a-c) shows the Random Forest classifications of altered and unaltered specimens using the entire geochemical dataset as a function of increasing minimum terminal node size (2^0, 2^4, and 2^8),

visualised on the feature-selected embedding. The embedding allows the visualisation of precisely which geochemical region is being affected by the regularisation. Figure 10(d) shows the out-of-bag error rates for the Random Forest, which increases monotonically as the forest is regularised.

The full-depth Random Forest (minimum terminal node size of 1) achieves excellent accuracy on the out-of-bag samples, however outliers are visible towards the bottom of the main cluster which are logged as metaperidotite but are geochemically similar to olivine pyroxenite. As the regularisation increases (minimum terminal node size of 16), these points are re-logged (Figure 10(b)). It becomes clear when the forest is too heavily regularised as large, homogeneously logged clusters of specimens are re-classified (Figure 10(c), red circle). Therefore, the adequate level of regularisation for this model is a minimum terminal node size of approximately 16 and certainly below 512. In addition, it should be noted that Random Forests are not based on pairwise distances as t-SNE embeddings are; the intuitive correspondence between the Random Forest classifications of the specimens and their location on the t-SNE embedding, for varying degrees of regularisation, validates the embedding.

4 CONCLUSIONS

In this study, two t-SNE embeddings of geochemical data from the Kevitsa Ni-Cu-PGE deposit (Lapland, Finland) were created to visualise geological patterns: one using all 31 elements, the other using a subset of eleven elements empirically determined (using a feature selection process) to differentiate between altered and unaltered specimens. The first embedding revealed that highly bimodal elements tended to control the distance-based cluster structure of the data, but did not adequately separate altered and unaltered specimens. The second embedding could lay out non-hydrated and hydrated host rocks (i.e., olivine pyroxenite and metaperidotite) such that the gradation of the former to the latter was apparent, which was not present in the embedding based on all elements. The second embedding was also demonstrated as an effective tool for interpreting the output of black box classifiers: visualising the altered/unaltered classifications of a Random Forest classifier trained at different levels of regularisation facilitated judgement of the level of regularisation, which aided in the re-logging of alteration status. Overall, the findings in this study

illustrate that t-SNE is capable of producing geochemical embeddings wherein clusters or intra-cluster structure may reflect external geological properties, however, this is reliant on exclusion of input elements which are statistically irrelevant to the desired geological property.

Future development that will increase the practical value of t-SNE embeddings is their integration with other geoscientific data, primarily spatial coordinates and petrophysical measurements. Incorporating spatial coordinates may produce spatially consistent clusters which are valuable for geochemical domaining (e.g., Le Vaillant et al., 2017). However, this approach eliminates spatial plotting as a form of independent validation (Templ et al., 2008). Incorporating petrophysical measurements could further separate lithologies: P-wave velocity and density jointly relate to lithology and degree of alteration in Kevitsa, and seismic reflection has been used to identify lithological contacts (Koivisto et al., 2015). These additional geoscientific data could be integrated with geochemical data directly alongside the ilr-transformed assays, but numeric scaling would be required so that their contribution to the inter-point distance is known relative to that from the elemental concentrations. Alternatively, separate pairwise conditional probabilities p_{ij} could be calculated for each type of geoscientific data, and subsequently combined in a weighted sum. This falls under the paradigm of multi-view learning, and indeed a multi-view variant of t-SNE has been proposed by Xie et al. (2011).

ACKNOWLEDGEMENTS

We wish to acknowledge First Quantum Minerals Ltd. for providing the geochemical dataset. This work was supported by the Robert and Maude Gledden Postgraduate Research Scholarship and First Quantum Minerals Ltd.

COMPUTER CODE AVAILABILITY

The R programming language scripts developed for this paper have been made available in a public repository (http://github.com/tom-a-horrocks/t-SNE-geochemistry). All queries can be directed to the first author.

REFERENCES

Aitchison, J., 1992. On criteria for measures of compositional difference. Mathematical Geology 24, 365–379. https://doi.org/10.1007/BF00891269

Aitchison, J., 1986. The statistical analysis of compositional data, Monographs on Statistics and Applied Probability. Chapman & Hall, Ltd., London.

Aitchison, J., 1984. Reducing the dimensionality of compositional data sets. Mathematical Geology 16, 617–635. https://doi.org/10.1007/BF01029321

Alahakoon, D., Halgamuge, S.K., Srinivasan, B., 2000. Dynamic self-organizing maps with controlled growth for knowledge discovery. IEEE Transactions on Neural Networks 11, 601–614. https://doi.org/10.1109/72.846732

Balamurali, M., Melkumyan, A., 2016. t-SNE based visualisation and clustering of geological domain, in: Neural Information Processing: 23rd International Conference, ICONIP 2016: Proceedings. Kyoto, Kansai, pp. 565–572. https://doi.org/10.1007/978-3-319-46681-1_67

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., 1984. Classification and regression trees, 1st ed, The Wadsworth Statistics/Probability Series. Wadsworth International Group, New York.

Chang, W.-C., 1983. On using principal components before separating a mixture of two multivariate normal distributions. Journal of the Royal Statistical Society: Series C (Applied Statistics) 32, 267–275. https://doi.org/10.2307/2347949

Cracknell, M.J., Reading, A.M., McNeill, A.W., 2014. Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests™ and Self-Organising Maps. Australian Journal of Earth Sciences 61, 287–304. https://doi.org/10.1080/08120099.2014.858081

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology 35, 279–300. https://doi.org/10.1023/A:1023818214614

Ellefsen, K.J., Smith, D.B., Horton, J.D., 2014. A modified procedure for mixture-model clustering of regional geochemical data. Applied Geochemistry 51, 315–326. https://doi.org/10.1016/j.apgeochem.2014.10.011

Fraser, S., Dickson, B., Kowalczyk, P., Sparks, G., 2005. And now for "SOM" thing completely different: spatial data mining, in: Window to the World: 2005 Symposium Proceedings. Geological Society of Nevada, Reno/Sparks, Nevada, p. 1310.

Fraser, S.J., Dickson, B.L., 2005. Ordered vector quantization for the integrated analysis of geochemical and geoscientific data sets, in: 22nd International Geochemical Exploration Symposium 2005: From Tropics to Tundra: Program and Abstracts. Association of Applied Geochemists, Perth, Western Australia, pp. 52–53.

Friedman, J.H., Tukey, J.W., 1974. A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers C–23, 881–890. https://doi.org/10.1109/T-C.1974.224051

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. Statistics and Computing 27, 659–678. https://doi.org/10.1007/s11222-016-9646-1

Gregory, J., Journet, N., White, G., Lappalainen, M., 2011. Technical report for the mineral resources and reserves of the Kevitsa project (Technical Report No. Ni 43-101). First Quantum Minerals, Ltd., Vancouver, British Columbia.

Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochemistry: Exploration, Environment, Analysis 10, 27–74. https://doi.org/10.1144/1467-7873/09-210

Gulson, B., Korsch, M., Dickson, B., Cohen, D., Mizon, K., Davis, J.M., 2007. Comparison of lead isotopes with source apportionment models, including SOM, for air particulates. Science of the Total Environment 381, 169–179. https://doi.org/10.1016/j.scitotenv.2007.03.018

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 417–441. https://doi.org/10.1037/h0071325

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Introduction, in: Haykin, S. (Ed.), Independent Component Analysis. John Wiley & Sons, Inc., New York, pp. 1–12.

Juntunen, P., Liukkonen, M., Lehtola, M., Hiltunen, Y., 2013. Cluster analysis by self-organizing maps: an application to the modelling of water quality in a treatment process. Applied Soft Computing 13, 3191–3196. https://doi.org/10.1016/j.asoc.2013.01.027

Kohonen, T., 1990. The self-organizing map. Proceedings of the IEEE 78, 1464–1480. https://doi.org/10.1109/5.58325

Koivisto, E., Malehmir, A., Hellqvist, N., Voipio, T., Wijns, C., 2015. Building a 3D model of lithological contacts and near-mine structures in the Kevitsa mining and exploration site, Northern Finland: constraints from 2D and 3D reflection seismic data. Geophysical Prospecting 63, 754–773. https://doi.org/10.1111/1365-2478.12252

Krijthe, J.H., 2015. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation.

Kyser, K., Barr, J., Ihlenfeld, C., 2015. Applied geochemistry in mineral exploration and mining. Elements 11, 241–246. https://doi.org/10.2113/gselements.11.4.241

Lacassie, J.P., Roser, B., Ruiz Del Solar, J., Hervé, F., 2004. Discovering geochemical patterns using self-organizing neural networks: a new perspective for sedimentary provenance analysis. Sedimentary Geology 165, 175–191. https://doi.org/10.1016/j.sedgeo.2003.12.001

Lacassie, J.P., Ruiz-Del-Solar, J., 2006. Knowledge extraction in geochemical data by using self-organizing maps, in: The 2006 IEEE International Joint Conference on Neural Network Proceedings. Institute of Electrical and Electronics Engineers, Vancouver, British Columbia, pp. 4878–4883. https://doi.org/10.1109/IJCNN.2006.247167

Le Vaillant, M., 2014. Hydrothermal remobilisation of base metals and platinum group elements around komatiite-hosted nickel-sulphide deposits: applications to exploration methods (PhD Thesis). University of Western Australia, Crawley, Western Australia.

Le Vaillant, M., Barnes, S.J., Fiorentini, M.L., Santaguida, F., Törmänen, T., 2016. Effects of hydrous alteration on the distribution of base metals and platinum group elements within the Kevitsa magmatic nickel sulphide deposit. Ore Geology Reviews 72, 128–148. https://doi.org/10.1016/j.oregeorev.2015.06.002

Le Vaillant, M., Hill, J., Barnes, S.J., 2017. Simplifying drill-hole domains for 3D geochemical modelling: An example from the Kevitsa Ni-Cu-(PGE) deposit. Ore Geology Reviews. https://doi.org/10.1016/j.oregeorev.2017.05.020

Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J., 2012. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. Computational Statistics & Data Analysis 56, 2688–2704. https://doi.org/10.1016/j.csda.2012.02.012

Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A., 2011. Dealing with zeros, in: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis. John Wiley & Sons, Ltd., Chichester, pp. 43–58. https://doi.org/10.1002/9781119976462.ch4

Meng, H.-D., Song, Y.-C., Song, F.-Y., Shen, H.-T., 2011. Research and application of cluster and association analysis in geochemical data processing. Computational Geosciences 15, 87–98. https://doi.org/10.1007/s10596-010-9199-x

Mutanen, T., 1997. Geology and ore petrology of the Akanvaara and Koitelainen mafic layered intrusions and the Keivitsa-Satovaara layered complex, northern Finland, Geological Survey of Finland, Bulletin. Geological Survey of Finland, Espoo.

Mutanen, T., Huhma, H., 2001. U-Pb geochronology of the Koitelainen, Akanvaara and Keivitsa layered intrusions and related rocks, in: Vaasjoki, M. (Ed.), Radiometric Age Determinations from Finnish Lapland and Their Bearing on the Timing of Precambrian Volcano-Sedimentary Sequences, Special Papers. Geological Survey of Finland, Espoo, pp. 229–246.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572. https://doi.org/10.1080/14786440109462720

Penn, B.S., 2005. Using self-organizing maps to visualize high-dimensional data. Computers & Geosciences 31, 531–544. https://doi.org/10.1016/j.cageo.2004.10.009

Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Cluster analysis, in: Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons, Ltd., Chichester, pp. 233–247.

Sammon, J.W., Jr., 1969. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers C–18, 401–409. https://doi.org/10.1109/T-C.1969.222678

Sanguinetti, G., 2008. Dimensionality reduction of clustered data sets. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 535–540. https://doi.org/10.1109/TPAMI.2007.70819

Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: problems and possibilities. Applied Geochemistry 23, 2198–2213. https://doi.org/10.1016/j.apgeochem.2008.03.004

Templ, M., Hron, K., Filzmoser, P., 2011. robCompositions: an R-package for robust statistical analysis of compositional data, in: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis. John Wiley & Sons, Ltd., Chichester, pp. 341–355. https://doi.org/10.1002/9781119976462.ch25

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–419. https://doi.org/10.1007/BF02288916

van der Maaten, L., 2014. Accelerating t-SNE using tree-based algorithms. Journal of Machine Learning Research 15, 3221–3245.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

van der Maaten, L., Postma, E., van den Herik, J., 2009. Dimensionality reduction: a comparative review (Technical Report No. TiCC TR 2009-005). Tilburg Centre for Creative Computing, Tilburg University, Tilburg, North Brabant.

Wattenberg, M., Viégas, F., Johnson, I., 2016. How to use t-SNE effectively. Distill. https://doi.org/10.23915/distill.00002

Xie, B., Mu, Y., Tao, D., Huang, K., 2011. m-SNE: multiview stochastic neighbor embedding. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41, 1088–1096. https://doi.org/10.1109/TSMCB.2011.2106208

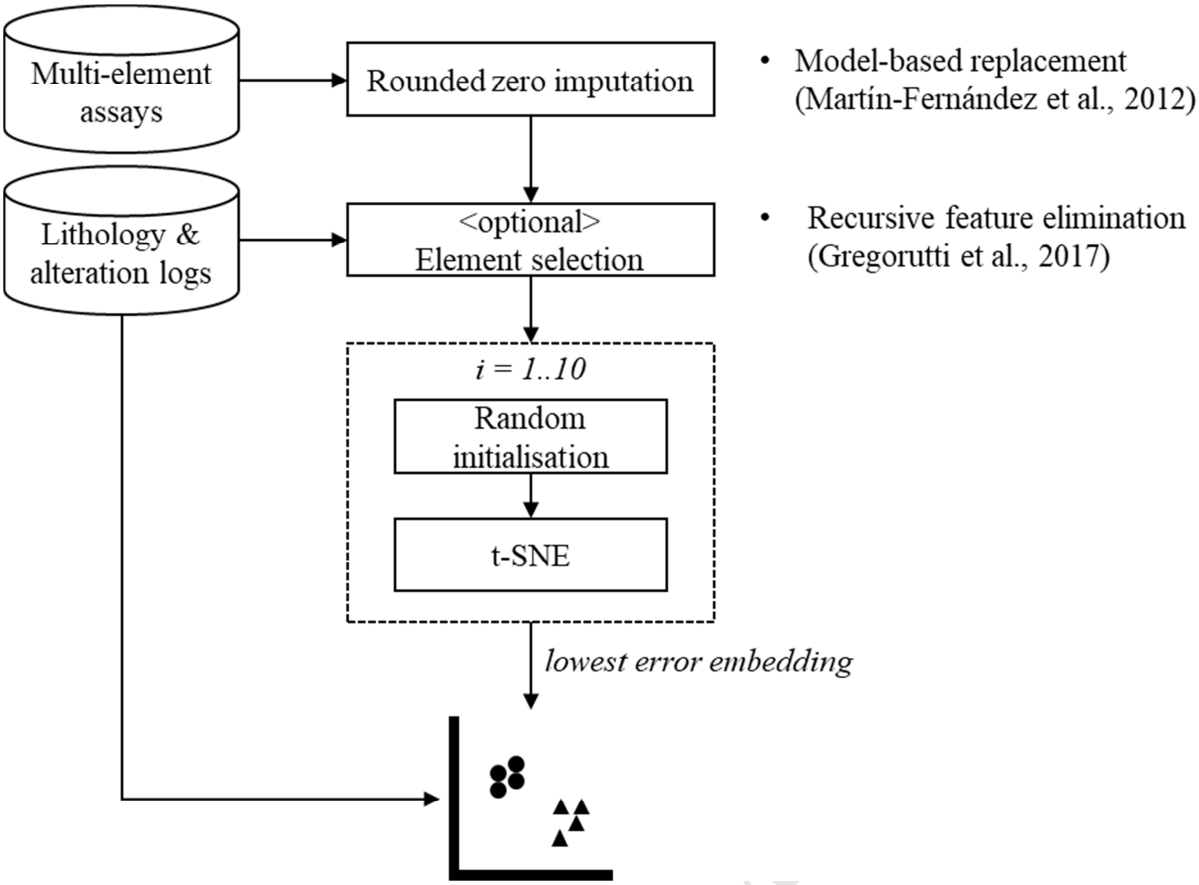


Figure 1 Overview of the dimensionality reduction technique; see Method (Section 2.2) for elaboration. Note that lithology and alteration logs may be used for element selection but are otherwise held out from the dimensionality reduction process.

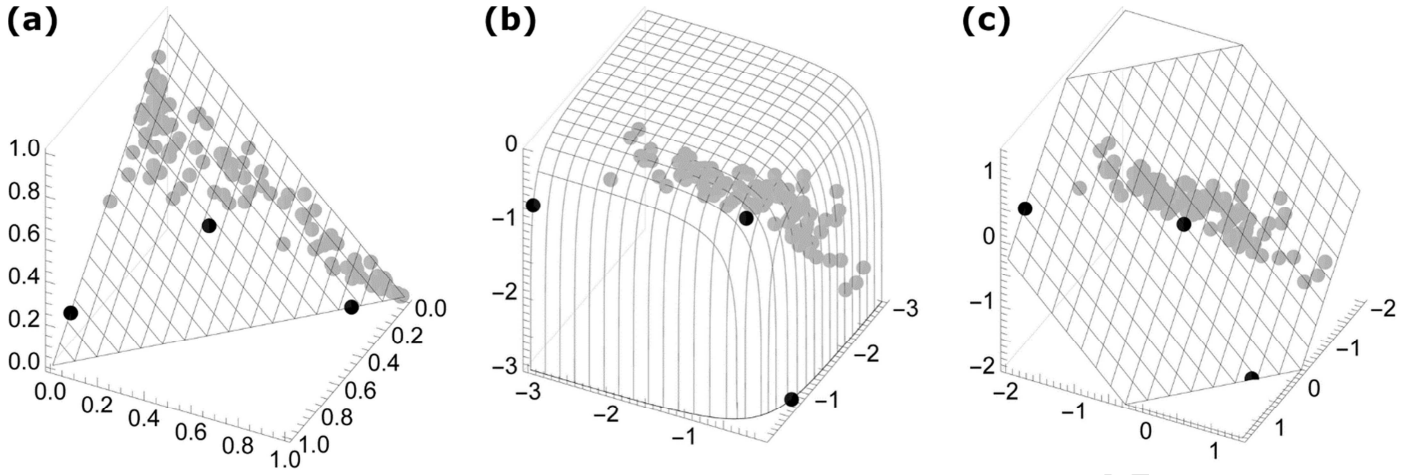


Figure 2 A synthetic compositional dataset in (a) regular space, (b) $\log_{10}$ space, and (c) $\log_{10}$ space with points normalised by their geometric mean. The straight-line distances in each subfigure correspond to (a) Euclidean distance, (b) Euclidean distance of log-transformed points, and (c) Aitchison distance. Note that the bolded points on the boundary are equidistant in (b) and (c), but are both closer to the bolded centre point in (c) since the Aitchison distance corrects for geometric mean.

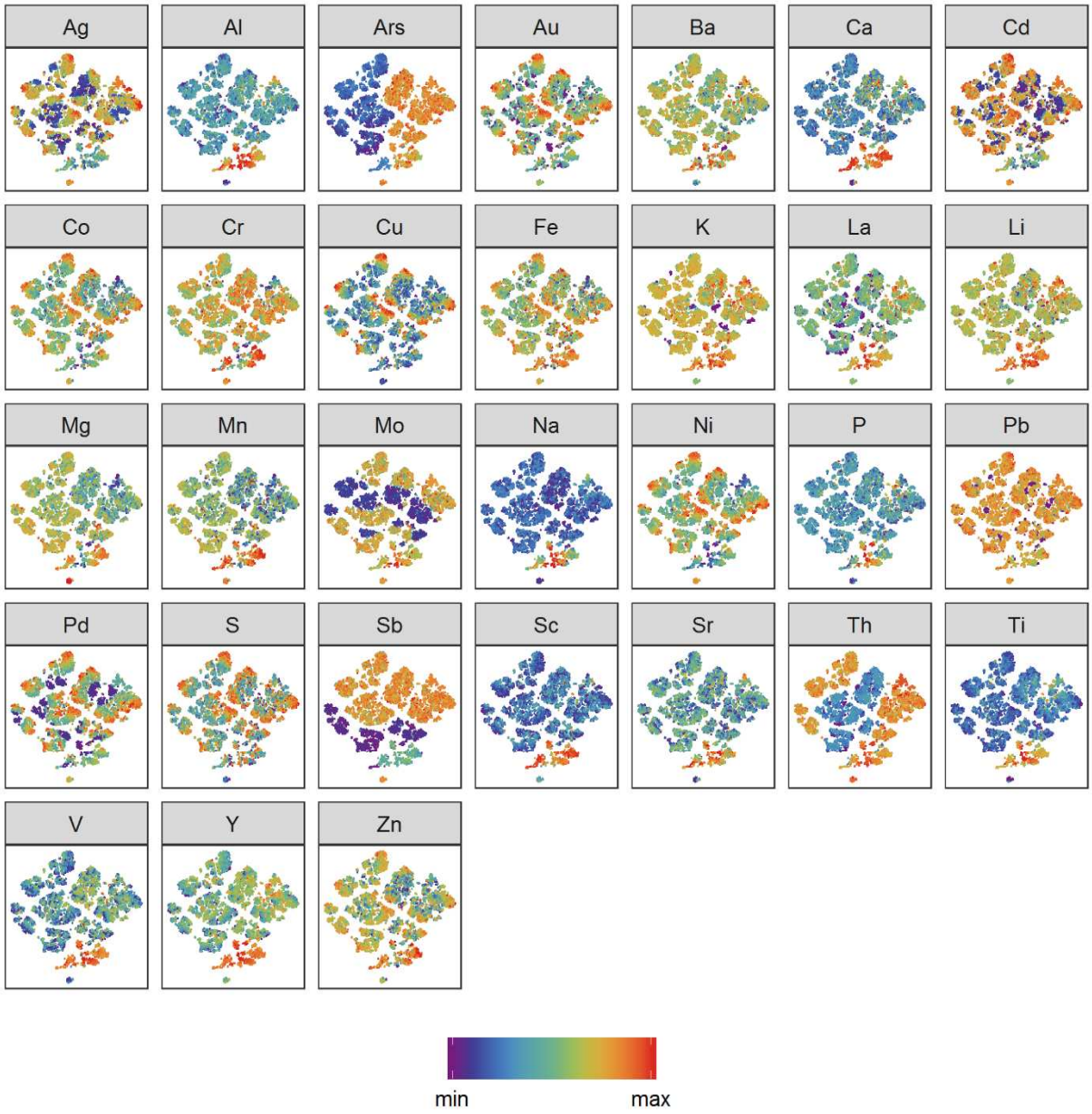


Figure 3 Embedding of the geochemical dataset with all elements, coloured by elemental concentrations (log-scale). Large-scale clustering is strongly influenced by elements with bimodal distributions such as arsenic and thorium, while elements with little influence appear in all clusters as peppered or in rainbow bands (e.g., gold and sulphur).

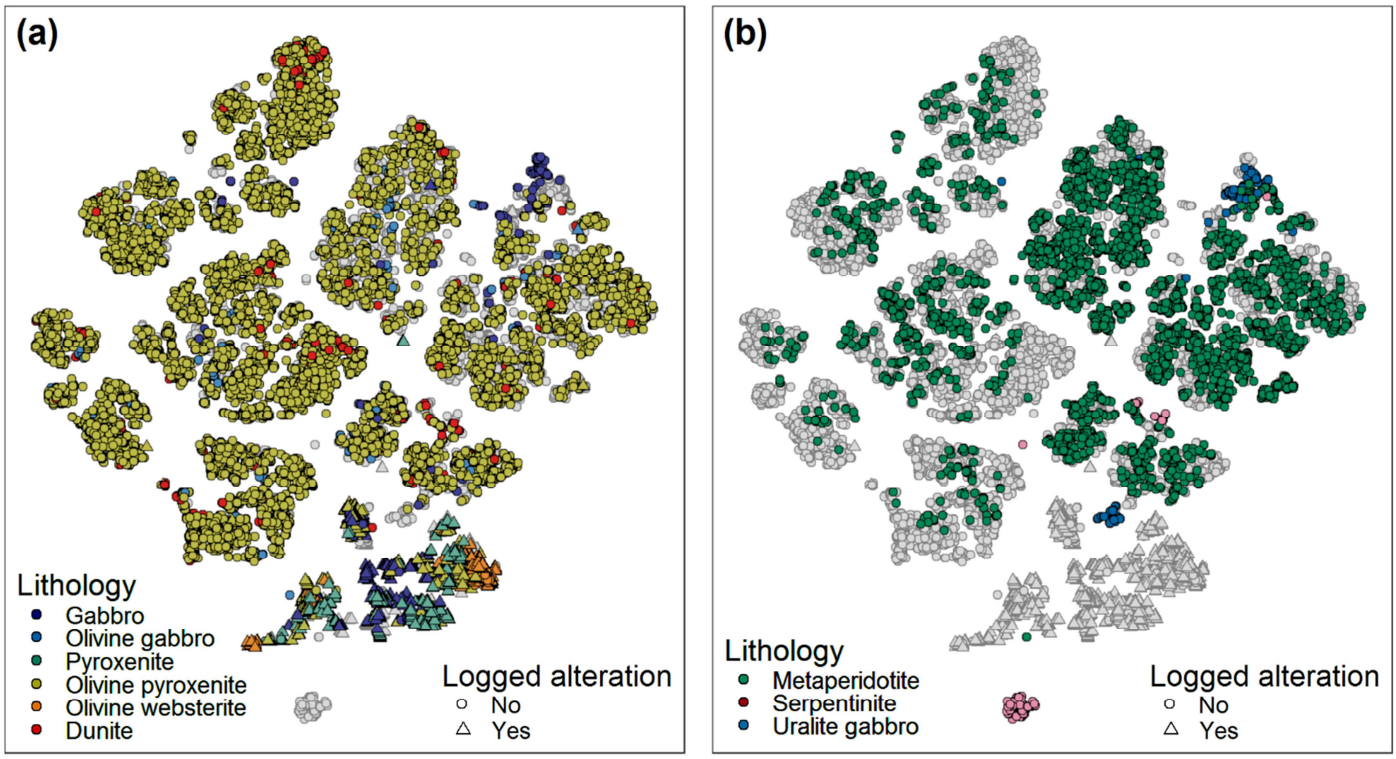


Figure 4 Embedding of the geochemical dataset of all elements, coloured by (a) unaltered logged lithology; and (b) altered logged lithology with colour approximately indicates olivine content (blue is low). Specimens with lithologies outside of the legend (including 18 minor lithologies) are coloured grey.

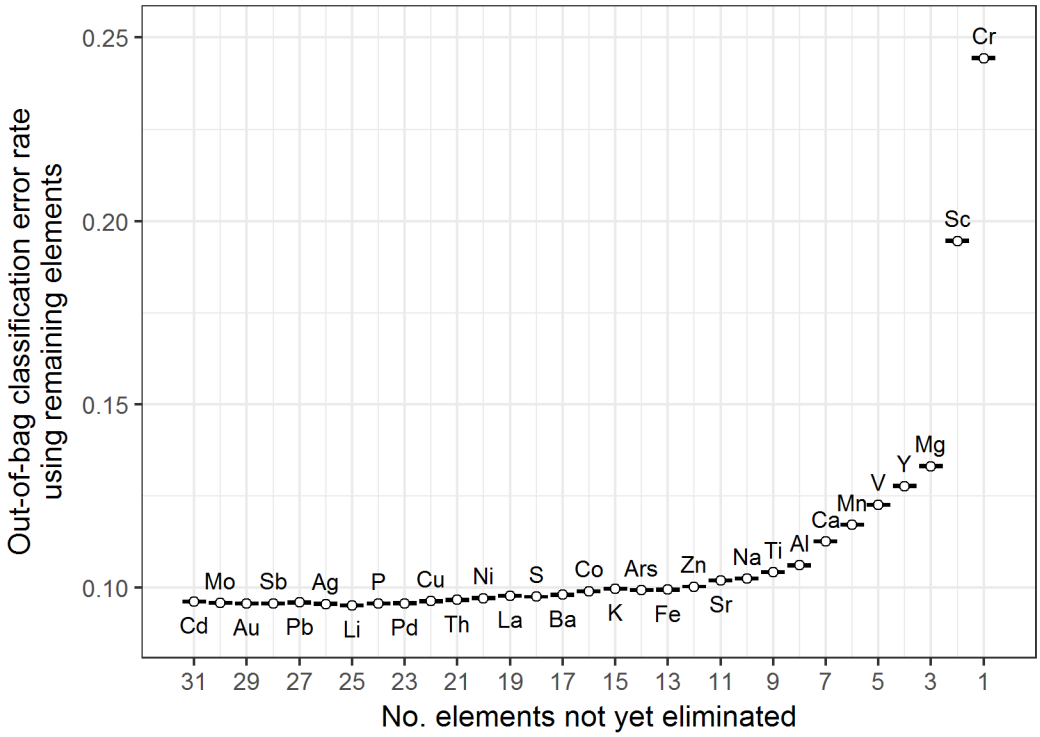


Figure 5 The Random Forest's out-of-bag classification error (y-axis) when discriminating between altered and unaltered specimens during recursive feature elimination. The x-axis shows how many elements are in the current iteration of the elimination procedure and the next element to be eliminated. Error bars represent the 95% confidence level in out-of-bag classification error rate over 20 forests.

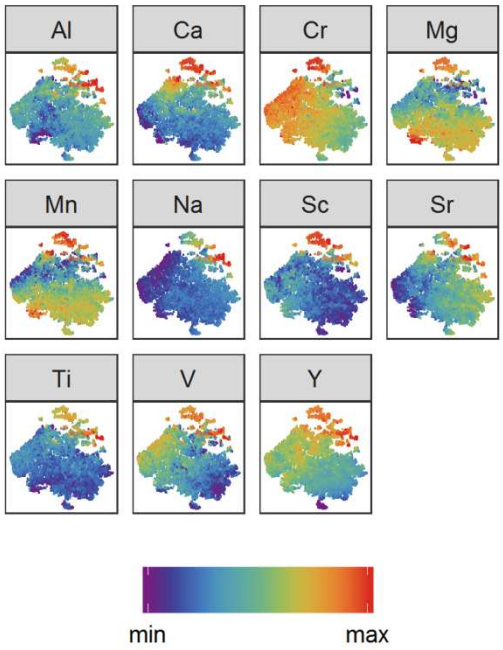


Figure 6 Embedding of selected elements from the geochemical dataset, coloured by elemental concentrations (log-scale).

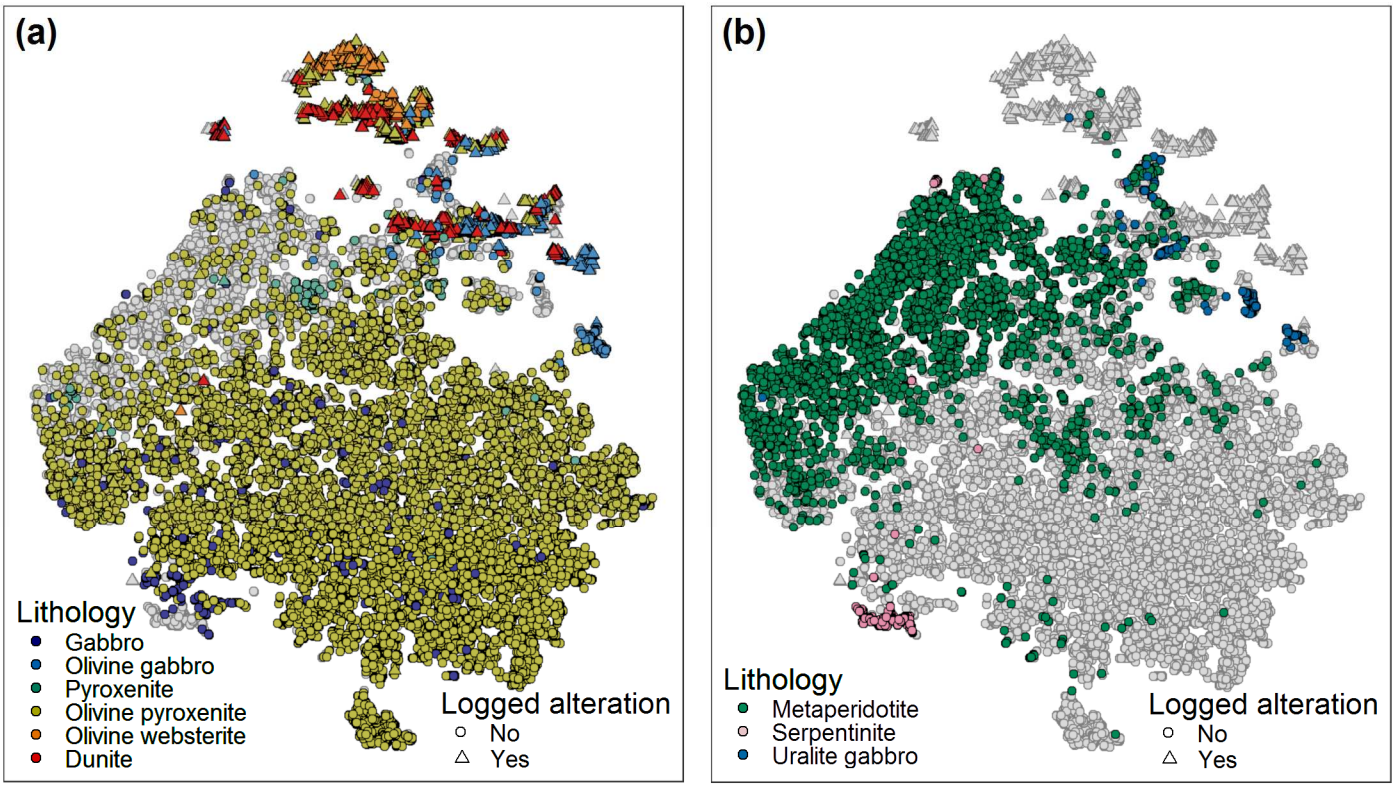


Figure 7 Embedding of selected elements from the geochemical dataset, coloured by (a) unaltered logged lithology; and (b) altered logged lithology. Refer to Figure 4's caption for colour scheme details.

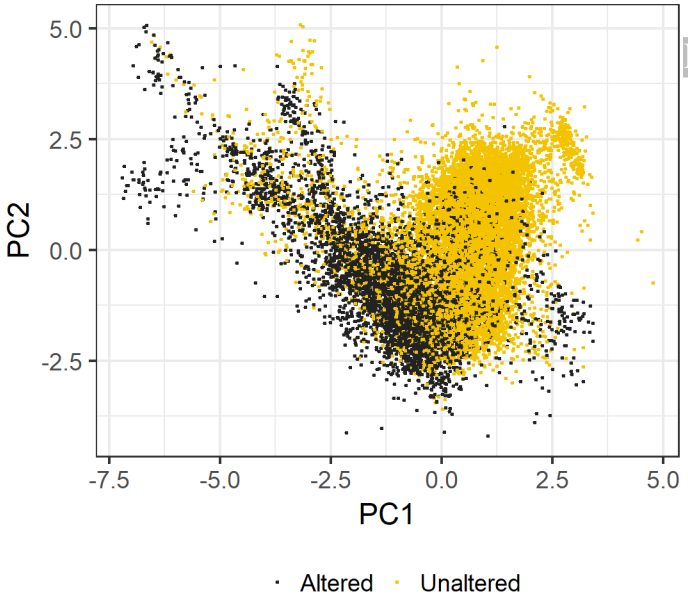


Figure 8 First two principal components generated from the geochemical dataset (selected elements only).

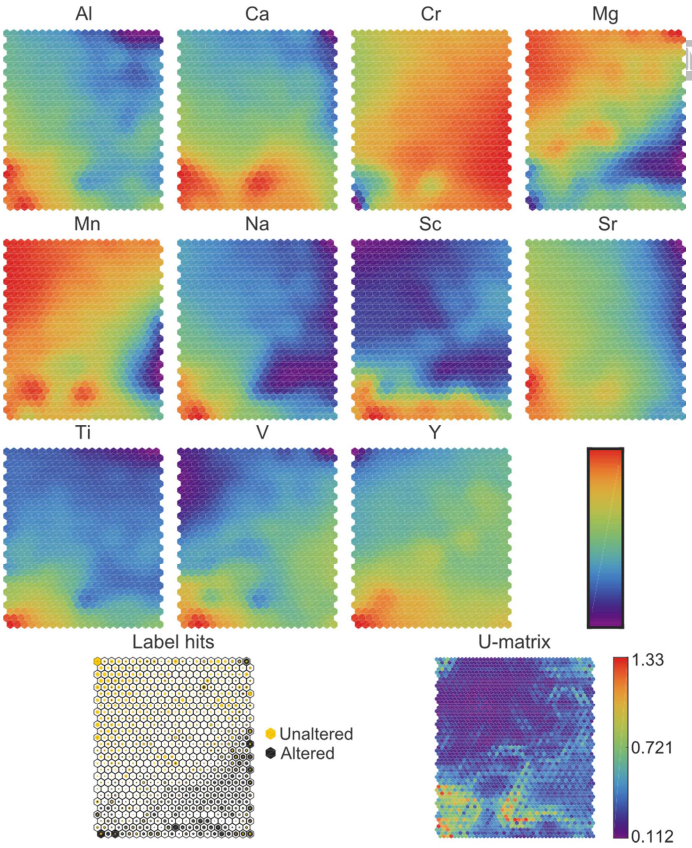


Figure 9 SOM embedding of the geochemical data subset of alteration discriminator elements. The top three rows show the log-scale elemental concentrations, while the bottom row shows: (left) the distribution of altered and unaltered specimens over the grid cells; and (right) the U-matrix, which outlines distinct regions in embedding.

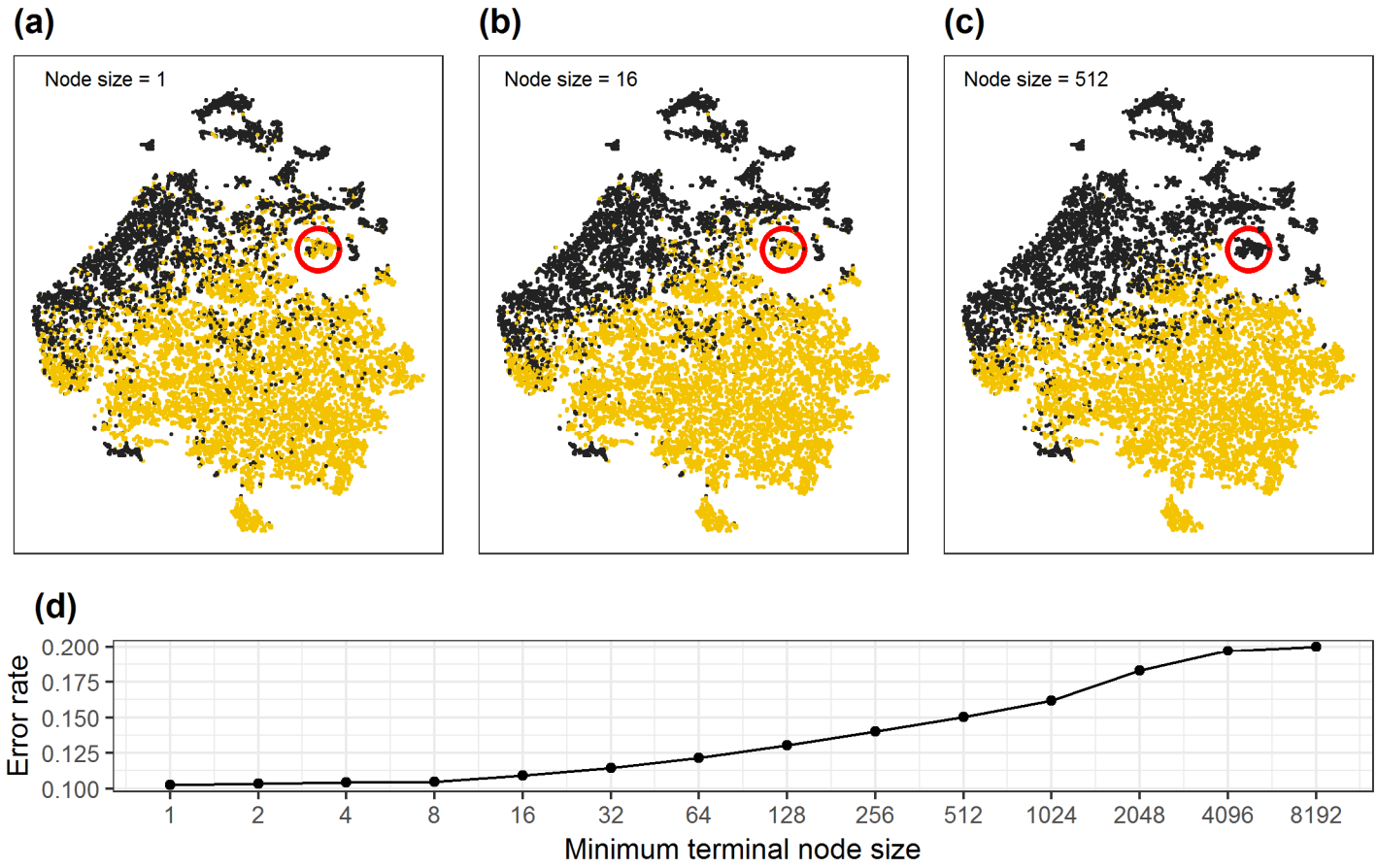


Figure 10 (a-c): Random Forest classifications of altered (black) and unaltered (yellow) specimens using the full geochemical dataset, but plotted on the alteration suite embedding, as a function of changing minimum terminal node size; (d): the classification error on out-of-bag specimens, which increases with minimum terminal node size. The embedding characterises how the Random Forest classifications change as it is increasingly regularised: the classifications in (b) are less peppered than in (a), signifying a simpler geochemical model, while (c) is simpler again at the expense of misclassifying a large cluster of specimens (red circle).

TABLES

Table 1 Population of specimens (drill core intervals) grouped by logged lithology.

Lithology	Count
Unaltered	**12 212**
Olivine pyroxenite	10 445
Gabbro	440
Pyroxenite	399
Dunite	276
Olivine websterite	227
Olivine gabbro	153
Diorite	86
Peridotite	55
Magnetite gabbro	36
Plagioclase bearing olivine websterite	36
Diabase	20
Granophyre	16
Pegmatite	9
Ultramafic (undifferentiated)	7
Microgabbro	3
Intrusive (mafic)	2
Intrusive (felsic)	1
Websterite	1
Altered	**3 953**
Metaperidotite	3 494
Serpentinite	159
Uralite gabbro	150
Hornfels	71
Completely altered (lithology unknown)	38
Albitite	21
Hornblendite	12
Meta-gabbro	7
Amphibolite	1

Table 2 Detection limits (DL) used for each element in the rounded zero imputation, and percentage of instances below detection limit (< DL). All concentrations were above detection limit for Al, Ca, Fe, Mg, Mn, Sr, and Ti.

	< DL (%)	DL	
Co	< 0.1	0.01	ppm
V	< 0.1	2.32	ppm
Cr	< 0.1	0.77	ppm
Sc	< 0.1	0.08	ppm
P	< 0.1	0.02	ppm
Na	< 0.1	0.01	%

Zn	0.3	0.20	ppm
Ba	0.6	0.04	ppm
S	0.7	0.01	%
Y	1.4	0.01	ppm
Li	1.4	0.01	ppm
Ni	1.8	0.01	%
K	2.8	0.36	ppm
Pb	3.7	0.01	ppm
La	5.7	0.01	ppm
Au	8.0	0.10	ppb
Cu	7.9	0.01	%
Pd	17	0.10	ppb
Sb	24	0.01	ppm
Cd	24	0.01	ppm
Ag	25	0.01	ppm
Mo	35	0.01	ppm
Th	41	0.01	ppm
Ars	50	0.01	ppm

Table 3 Minimum, maximum, and 1st and 99th percentile concentrations of all assayed elements selected for further analysis.

		Percentiles		
	Min.	1st	99th	Max.
Ag (ppm)	0.000 233	0.00150	3	8.61
Al (%)	0.007 73	0.0864	7.78	10.3
Ars (ppm)	6.30×10^{-6}	5.20×10^{-5}	40.9	3500
Au (ppb)	7.08×10^{-7}	0.0551	270	4700
Ba (ppm)	0.0350	1	200	754
Ca (%)	0.008 50	0.0941	10.1	39.9
Cd (ppm)	0.000 384	0.002 53	1.01	4.75
Co (ppm)	0.0100	15.9	237	2590
Cr (ppm)	0.770	6.82	1680	11 100
Cu (%)	3.72×10^{-8}	0.003 34	0.860	7.79
Fe (%)	0.348	1.25	11.2	60.8
K (%)	2.19×10^{-5}	2.78×10^{-5}	1.76	5
La (ppm)	0.003 81	0.006 78	33.1	493
Li (ppm)	0.007 48	0.008 65	17.4	80.3
Mg (%)	0.0219	0.496	19.6	27.2
Mn (ppm)	19	89	1470	4840
Mo (ppm)	0.000 275	0.001 07	3.51	184
Na (%)	0.007 72	0.0213	4.02	8.13
Ni (%)	1.48×10^{-6}	0.003 80	0.510	4.16
P (ppm)	0.0200	8.71	940	4640
Pb (ppm)	0.00360	0.007 26	16.9	50.1
Pd (ppb)	1.31×10^{-8}	0.0227	464	8090
S (%)	2.77×10^{-7}	0.0100	3.37	26.5
Sb (ppm)	0.000 54	0.001 36	21.8	43.8
Sc (ppm)	0.0618	0.780	56.2	84.5

Sr (ppm)	0.300	1.18	176	448
Th (ppm)	2.08×10^{-6}	4.18×10^{-5}	9.70	57.4
Ti (%)	0.000 125	0.008 82	0.723	1.69
V (ppm)	2.19	7.59	293	6290
Y (ppm)	0.007 30	0.008 84	22.4	63.9
Zn (ppm)	0.185	2.11	78	888

Table 4 Effect of element selection on nearest neighbour reclassification accuracy ('altered' vs. 'unaltered'), as applied to the original (ilr-transformed) data, the t-SNE embedding, and the first two principal components.

	Dims	**Accuracy (%)**	
All elements		altered	unaltered
Original (ilr)	30	75.5	87.7
t-SNE	2	65.5	82.4
PCA	2	54.2	75.9
Selected elements			
Original (ilr)	10	83.3	92.2
t-SNE	2	81.8	91.0
PCA	2	76.6	90.3

* Dimensionality reduction applied to Kevitsa's multi-element drill core assay database
* Produced representations separated hydrated core despite absence of loss on ignition
* Assay-based core re-logging visualised using representations
* Feature selection prior to dimensionality reduction improved produced representations
* Compositional nature of assays addressed using Aitchison distance