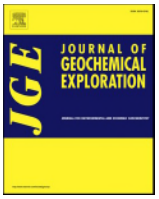
Journal of Geochemical Exploration 258 (2024) 107388

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

journal homepage: www.elsevier.com/locate/gexplo



Using machine learning to identify indicators of rare earth element enrichment in sedimentary strata with applications for metal prospectivity

Brendan A. Bishop *, Leslie J. Robbins

Department of Geology, University of Regina, 3737 Wascana Parkway, Regina, Saskatchewan S4S 0A2, Canada

ARTICLE INFO

Keywords:
Rare earth elements
Energy transition
Machine learning
Compositional data analysis
Economic geology

ABSTRACT

Rare earth elements (REE), classified as critical minerals which are crucial for clean energy technologies, face soaring demand. While economic deposits are found in limited geologic environments including carbonatites and ion-adsorption clays, unconventional, secondary sources such as those from sedimentary basins could hold potential to meet this increased demand. Coal and its associated combustion by-products, phosphorites, oil sands tailings, and formation waters have all garnered interest for REE recovery, yet they remain significantly underexplored. Accordingly, new tools for data analysis and optimization such as machine learning can assist in mineral prospectivity, with these tools being subject to rapid proliferation in the Earth sciences.

This work leverages compositional data analysis principles and machine learning to probe geochemical relationships and predict REE abundances in sedimentary lithologies using unsupervised (correlation, principal component, and cluster analysis) and supervised (regression, support vector machine, random forest, and boosting) machine learning models. These three unsupervised models display similar results, with REE typically being associated with incompatible elements (e.g., Th, Nb, and Hf). Gradient boosting, Adaboost, and Random Forest had the highest performance for predicting REE concentrations, with Th and P commonly being the most important predictor variables. Identifying geochemical indicators of REE enrichment that may be used to assist in discovering potentially exploitable REE resources based on existing data, as well as increasing the understanding of metal behaviour in sedimentary systems, is a step forward in understanding novel secondary and unconventional REE sources. Although REE concentrations from these sources are generally lower than primary ore deposits, the amount of available feedstock, potentially simpler, cheaper, and less environmentally taxing extraction processes, and the added benefit of remediating waste streams and contributing to the circular economy make these sources alluring.

1. Introduction

The transition to a low carbon economy will require a substantial increase in metal production (Lee et al., 2020), which, in turn, will result in mineral exploration expanding to deeper, more geologically complex, and previously unexplored settings. This will necessitate the need for new methods for analyzing and optimizing the use of data throughout the exploration process, such as the deployment of machine learning-based approaches (Caté et al., 2017). To this end, machine learning is being increasingly utilized in the Earth sciences due to improvements in computational resources, an ever-increasing amount of data, and the availability of publicly accessible geochemical and remote sensing datasets (e.g., Engle and Brunner, 2019; Karpatne et al., 2019). These datasets are amenable to analysis by machine learning algorithms (MLAs) as they are typically comprised of a large number of samples with many variables (Lindsay et al., 2021). Machine learning is a branch of artificial intelligence (AI) used to identify patterns within data and make predictions based on those patterns and includes both unsupervised and supervised methods (Caté et al., 2017). Commonly used MLAs consist of both supervised and unsupervised models which can be used to decipher complex relationships among variables (e.g., elements) and to complement a hypothesis-driven approach (Zhu et al., 2023). Such MLAs have been used to predict mineralization and elemental concentrations (Caté et al., 2017; Schnitzler et al., 2019; Grunsky and de Caritat, 2020), classify samples (Engle and Brunner, 2019; Gregory et al., 2019), and identify patterns and reduce dimensionality (Hu et al., 2022; Lindsay et al., 2021).

Rare earth elements (REE) are often classed as critical metals (e.g.

* Corresponding author.
E-mail address: bab495@uregina.ca (B.A. Bishop).

https://doi.org/10.1016/j.gexplo.2024.107388
Received 4 September 2023; Received in revised form 12 December 2023; Accepted 20 December 2023
Available online 8 January 2024
0375-6742/© 2024 Elsevier B.V. All rights reserved.

European Commission, 2020; Natural Resources Canada, 2022; US Geological Survey, 2022) and are essential for the transition to a low carbon economy, as they are integral in clean energy technologies. Therefore, it has been estimated that demand will double in the next ten years requiring a significant increase in global production: up to the equivalent of one Mount Weld or Mountain Pass deposit per year (Goode, 2023). While REE are not necessarily rare with respect to crustal abundances, economic concentrations are confined to few geologic environments (Linnen et al., 2014). The majority of REE production comes from carbonatite and ion adsorption clay deposits; however, these are scarce and until recently have not been the subject of significant exploration (Balaram, 2019). However, issues surrounding the development of major REE mining operations include long lead times and a variety of environmental concerns (Yin et al., 2021). Considering the above-mentioned variables and a desire to secure a reliable domestic REE supply chain, in part related to geopolitical concerns, there has been increasing interest toward extracting REE from unconventional, secondary sources (Dushyantha et al., 2020).

Sedimentary environments host several potential REE sources, including coal and coal waste (Creason et al., 2023), phosphorites (Emsbo et al., 2015), oil sands tailings (Roth et al., 2017), deep-sea muds (Kato et al., 2011), and geothermal and formation waters (Quillinan et al., 2018; Miranda et al., 2022). While these sources have not been thoroughly investigated for their REE potential, a significant volume of data has been produced through industrial activity, including fossil fuel exploration, making this data well suited to analysis by machine learning. Additionally, previous data analysis has indicated that REE, together with major and trace elements, may be important indicators for mineral exploration (Vural, 2020). For instance, Montross et al. (2022) developed a method for collecting and analyzing data, primarily from drill core, to predict REE concentrations in sedimentary strata with a focus on coal bearing intervals.

Inspired by the increasing demand for REE and proliferation of MLAs in Earth science, this study applies compositional data analysis and machine learning to investigate geochemical relationships and develop a model to predict REE abundances in sedimentary lithologies. Data for this study was acquired from open-source repositories including the Sedimentary Paleoenvironments Project (SGP) (Farrell et al., 2021), the Alberta Geological Survey, and the Saskatchewan Geological Survey. Compositional data analysis coupled with unsupervised MLAs including correlation analysis, principal component analysis (PCA), and cluster analysis were performed to determine the relationships between variables (i.e., elements), while supervised machine learning models were implemented to predict REE concentrations. While spatial data can also be important, this work only considered geochemical relationships and is meant to provide a starting point for assessing potential targets for the extraction of REE from unconventional sedimentary environments, as it can be used to gain a high-level understanding and predict whether a specific setting could contain economic REE abundances. While market conditions play a significant role in determining what would be considered an economic resource, understanding these sources and having the ability to predict enrichments is crucial for guiding future exploration and can be implemented using data that is readily available. Finally, findings gleaned here can also be applied to investigations focusing on sedimentary environments to better understand their geological history, making these tools valuable from both an economic and Earth systems perspective.

2. Theoretical background and methods

2.1. Data acquisition

Geochemical data for this study was compiled from the Saskatchewan Geological Survey (Jensen et al., 2020), the Alberta Geological Survey (Lopez et al., 2020; Rokosh et al., 2016), and the SGP database (Farrell et al., 2021). From these datasets, samples that contained lithology, geologic formation, and geologic period information as well as compositional data including major oxides, trace metals, and REE were selected. Where lithology information was not given but formation name exists, lithology was assigned based on the dominant lithology reported in the literature.

Geochemical data is commonly reported as a combination of oxide weight percent (wt%; for major elements) and parts per million (ppm; for trace elements). Since many multivariate techniques are based on distance coefficients, the variable with the greatest magnitude will have the greatest impact on the outcome, and therefore the units for each variable must be consistent (Templ et al., 2008). As such, all elemental concentrations are presented in ppm, which required converting oxide abundances in wt% to equivalent ppm values. Although strategies for treating missing data exist, imputation approaches can be complicated and time-consuming, while replacement or interpolation methods can introduce additional uncertainty by affecting covariance and correlation between variables (Hastie et al., 2009; Zhu et al., 2023). Therefore, the goal of data cleaning was to build a complete dataset where each sample contained major compositional elements (Al, Ca, Fe, K, Mg, Na, P, Si, and Ti) and REE as well as select trace elements with no missing values. For samples which were below the detection limit, their concentration was set as half the detection limit. No a priori QA/QC procedure was performed on the data to remove outliers and analytical errors; this ensures the datasets are as realistic as possible and do not represent an idealized starting point. Dataset 1 which included REE and sample context information (i.e. formation, locality, and lithology) contained 4364 observations (Table SI1), while the Dataset 2 used in the machine learning contained 3527 samples with 36 elements (Al, Ba, Ca, Co, Cs, Fe, Ga, Hf, K, Mg, Na, Nb, Ni, P, Rb, Si, Sr, Th, Ti, U, V, Zr, and REE) (Table SI2). Phosphorites were excluded from Dataset 2 and multivariate analyses due to several variables being absent. In this work, Y is considered a REE and included in the analysis. Although Sc is sometimes included as a REE, it exhibits different geochemical behaviour and is typically not found in deposits that host REE (Jowitt, 2022), and therefore was not included here. For the machine learning, the REE were summed and presented as a single variable. Data cleaning and analysis was performed using the R programming language (R Core Team, 2020).

2.2. Compositional data

Since geochemical data captures the relative proportions of elements within a sample (e.g., rock, mineral, water), it is considered compositional in nature: multivariate data where components represent part of a whole that sums to a constant (i.e., 1 or 100%) (Aitchison, 1986). While variables that comprise most data types are free to vary from $-\infty$ to $+\infty$, compositional data occupies a restricted space, referred to as the simplex, where variables can only vary between 0 and 100% (Egozcue et al., 2003). Since the data has a constant sum, it is considered to be constrained as the variables cannot vary independently, forcing at least one covariance in the data to be negative generating a bias toward negative correlation (Pawlowsky-Glahn and Egozcue, 2006). As a result, standard statistical techniques that are commonly used for unconstrained random variables, such as PCA, cannot be used to analyse compositional data in its raw form since it can lead to multiple issues, including results that have little geological significance (Pawlowsky-Glahn and Egozcue, 2006). Aitchison (1986) recognized that it was the relative magnitude and variation of components, as opposed to their absolute values, which are important for analyzing compositional data. Subsequently, approaches based on log-ratio transformations, including the additive-log-ratio (ALR), centred-log-ratio (CLR) (Aitchison, 1986), and more recently, isometric-log-ratio (ILR) (Egozcue et al., 2003) transformations have been developed. Although the ILR transformation is the most mathematically robust of these transformations, the dimensionality of the dataset is reduced by one and the variables are therefore no longer directly interpretable (Templ et al., 2008; Xie et al., 2018; Bern et al., 2021), accordingly, the CLR transformation was employed here.

Although implementing log-ratio transformations when analyzing compositional data may seem complicated and unnecessarily complex, past work has demonstrated that improper treatment of environmental and geochemical data can generate misleading or spurious results (Engle and Rowan, 2014). Here, CLR transformations were performed using the compositions package V2.0-2 in R (Van Den Boogaart and Tolosana-Delgado, 2008).

2.3. Unsupervised machine learning

Correlation analysis, PCA, and cluster analysis were used to identify relationships between variables in the data and, since they each use different mathematical methods, similarities and differences between each algorithm were assessed. Correlation analysis is an important exploratory data analysis tool as it provides a quantitative method for determining whether variables are related (Filzmoser and Hron, 2009). Here, correlation matrices with Pearson's correlation coefficients were created using the "corrplot" package (Wei and Simko, 2021) on both the untransformed and CLR transformed geochemical data.

Both PCA and cluster analysis are unsupervised machine learning techniques used in exploratory data analysis to reveal structure within a dataset. PCA is a multivariate data analysis procedure used to reduce the dimensionality of a dataset that consists of several interrelated variables (Jolliffe, 2002). It transforms the dataset into a new set of Principal Components (PCs), in which the initial components assume most of the variation from the original variables and the first PC captures the maximum variance (Jolliffe, 2002). By discarding the latter PCs with lesser variance, the method is commonly used for dimensionality reduction. PCA is typically shown using biplots, where the axes are the principal components selected for visualization. In a biplot of CLR-transformed data, the correlation coefficient of two variables is approximated by the cosine of the angle between two rays; accordingly, if two rays are near each other the corresponding variables may be highly correlated (Otero et al., 2005). PCA has been previously applied to geochemical data, for example, Lindsay et al. (2021) performed a PCA on elemental data from basaltic lava flows to determine how elemental concentrations were related and which elements were controlled by similar factors. Similarly, Bhuiyan et al. (2019) used PCA to establish geometallurgical relationships for a Brazilian gold mine. Other examples of PCA used in geochemical studies include Bishop et al. (2023) who used this approach to probe elemental associations in coal combustion by-products, and Mänd et al. (2021) who employed it to identify metal associations and sources in Paleoproterozoic chemical sediments. Since PCA is not designed for compositional data, applying the method directly to geochemical data can yield misleading results (Filzmoser et al., 2009), therefore the data was CLR transformed prior to PCA and visualized using the factoextra R package (Kassambara and Mundt, 2020).

Cluster analysis is a mathematical distance-based algorithm used to partition multivariate observations into several homogeneous groups where the observations are mapped into centroids (Templ et al., 2008). This partitions the dataset based on similarities between variables and summarises the data allowing for a better overview of its structure (Templ et al., 2008). The ideal outcome results in clusters where the samples within a cluster are as similar as possible, while the distances between clusters are as large as possible (Reimann et al., 2008). For geochemical data, cluster analysis can be utilized to detect relationships between variables (R-mode) or assign samples to specified classes or subsets for further analysis (Q-mode) (Templ et al., 2008). Cluster analysis was used by Ahmed et al. (2020) on LA-ICP-MS data to identify epidote samples with similar chemistries within and between samples, while Zhou et al. (2018) utilized cluster analysis to explore for potential Au mineralization within a deposit. Several clustering algorithms have been developed including hierarchical, partitioning, model-based, and fuzzy methods (Reimann et al., 2008). Hierarchical clustering is an agglomerative method which combines observations into clusters with pairs of clusters being merged as the hierarchy increases; while partitioning methods, including k-means, classify observations into groups and require the number of clusters to be pre-determined (Reimann et al., 2008). Here, hierarchical Q-mode cluster analysis was performed using the robCompositions R package (Templ et al., 2011) on the CLR transformed data.

2.4. Development and overview of the supervised machine learning models used in this study

A multitude of supervised MLAs have been implemented to tackle both regression and classification problems within the realm of geochemistry (Rodriguez-Galiano et al., 2015; Xie et al., 2018). In this study, Linear Regression (LR), Ridge Regression (RR), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Random Forest (RF), AdaBoost (Adaptive Boosting; AB), and Gradient Boosting (GB) were implemented to predict REE concentrations and to identify the models which have the highest performance.

The LR model is the simplest MLA which estimates the probability for a given feature and label directly from the training data (Ogen et al., 2022), where the objective is to find the plane that minimizes the sum-of-squared error (SSE) between the observed and predicted responses (Kuhn and Johnson, 2013). RR is a shrinkage model that adds a penalty on the sum of the squared regression parameters, developed by Hoerl and Kennard (1970). The penalty is added to the sum of the square regression parameters, where the parameter estimates are only allowed to become large if there is a proportional reduction in SSE which shrinks the estimates toward zero as the penalty increases (Kuhn and Johnson, 2013). Although adding the penalty increases the bias, it simultaneously reduces the variance enough to make the error smaller than unbiased models (Kuhn and Johnson, 2013). To predict mineralization in the Bathurst Mining Camp, New Brunswick, RR, in addition to Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net Regularized regression, were employed to overcome the challenges of weak geochemical and geophysical signals, over-fitting, and uncertainty of previous predictive models, with RR performing the best (Parsa et al., 2022).

KNN is a nonparametric method and among the simplest MLAs, first developed by Fix and Hodges (1951) and expanded on by Cover and Hart (1967). This method predicts a new sample using the K-closest samples from the training set by identifying that sample's nearest neighbours in the predictor space where the predicted response is the mean of the K neighbours' responses (Kuhn and Johnson, 2013). While this MLA performs well without needing adjustments, population size can significantly slow execution speed (Song et al., 2017). Kaplan and Topal (2020) found KNN in conjunction with Artificial Neural Networks (ANN) were suitable to perform gold grade estimations utilizing geochemical, alteration, and spatial data.

The SVM algorithm was proposed by Boser et al. (1992) and is based on maximizing the margin between training patterns and the decision boundary. It creates a set of hyperplanes to distinguish between instances of different classes and seeks to find the optimal hyperplane that maximizes the margin between the different classes (Xie et al., 2018; Bern et al., 2021; Chen et al., 2023). It has been applied to map prospectivity for gold deposits in Nova Scotia, Canada (Zuo and Carranza, 2011), classify mudstone lithofacies in the complex depositional environments of the Bakken and Marcellus shales of North America (Bhattacharya et al., 2016), and detect gold mineralization-related geochemical anomalies in Hebei Province, China (Chen et al., 2023).

The RF, AB, and GB algorithms are ensemble methods used to improve the predictive accuracy of decision trees. RF was developed by Breiman (2001) and utilizes a combination of bagging and the Classification and Regression Trees (CART)-split criterion (Breiman, 2001; Biau and Scornet, 2016;). Bagging (bootstrap-aggregating) is a general aggregation scheme where the original dataset is resampled randomly without replacement and is among the most effective prediction

techniques for large, high-dimension datasets (Rodriguez-Galiano et al., 2015; Biau and Scornet, 2016). This allows the user to build optimal decision trees by combining multiple iterative trees built from randomly selected samples (Schnitzler et al., 2019). AB and GB both use boosting, a process used to improve the performance of an MLA that combines the outputs of many "weak" classifiers to produce a powerful "committee" (Hastie et al., 2009). AB was developed by Freund and Schapire (1996) and is a model that iteratively trains a series of weak classifiers on different weighted subsets of the data and assigns higher weights to misclassified instances to create a stronger model by combining their predictions. Similarly, GB consecutively fits new models based on the weak learners to improve the accuracy of the response variable, where new base learners are constructed to be maximally correlated with the negative gradient of the loss function (Natekin and Knoll, 2013). In recent years, MLAs based on boosting and decision trees have been among the most popular MLAs in mineral prospectivity. This approach has been used to predict Na concentrations based on data from Quebec, Canada that included density, magnetic susceptibility, geochemical elements, average visible light reflectance, and infrared spectrometry, since Na is typically missing or difficult to measure by handheld X-ray fluorescence (Schnitzler et al., 2019). However, it is a crucial element for characterizing hydrothermal alteration in volcanogenic massive sulphide (VMS) deposits and is therefore useful in mining exploration (Schnitzler et al., 2019). Similarly, Gregory et al. (2019) used RF to classify the composition of pyrites by deposit type, which allows for the early implementation of predictive ore deposit models when prospecting in greenfield terrains. Lawley et al. (2022) used numerous MLAs including GB, RF, and Extreme Gradient Boosting (XGBoost) on geoscience datasets from Australia, Canada, and the USA to train prospectivity models for Mississippi Valley Type (MVT) and Clastic Dominated (CD) deposits at the continental scale.

Here, machine learning was implemented using the open-source data mining program Orange (Demsar et al., 2013). Parameters for each model are presented in Appendix 1. Ten-fold cross validation was used to evaluate the performance of the supervised machine learning models. This technique reduces overfitting and involves splitting the data into 10 equally sized *k* folds, where *k*-1 folds are used to build the model and the remaining fold is used for validation (Hastie et al., 2009). The folds were randomly assigned from the dataset, and the training and testing was repeated ten times, once for each fold with the average accuracy of the ten being taken as the model predictive accuracy. Model accuracy was assessed using the root mean square error (RMSE) as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}}$$

where $\widehat{y}_i$ and y_i are the predicted and true values of the *i*-th samples, respectively, and n is the number of samples. The models with the highest accuracy have the lowest RMSE values.

3. Results

3.1. Exploratory data analysis and unsupervised machine learning

A histogram of total REE concentrations is presented in Fig. 1, showing the data was strongly right skewed and hence does not follow a normal distribution. Of the 4364 data points in Dataset 1, only 69 points were above 500 ppm and 15 points were above 1000 ppm. The median and mean REE concentrations were 136.6 ppm and 157.4 ppm, respectively, which was slightly lower than 168.4 ppm for the Upper Continental Crust (UCC; Taylor and McLennan, 1985) and 173 ppm for the North American Shale Composite (Gromet et al., 1984).

Summary statistics for REE from Dataset 1 are outlined in Table 1. Clastic lithologies dominated the dataset at ~86% of all samples, carbonate lithologies encompassed 11%, and chemical and phosphorites comprised the remainder. Phosphorites were classified separately due to their anomalously high REE abundances and because they were from a single formation, the Phosphoria Formation in the western United States. Based on the information available from the compiled datasets, samples were further subdivided into individual lithologies, with shales making up the majority of clastics, followed by mudstones, siltstones, and sandstones. Limestone was the dominant carbonate lithology. There were a high number of undifferentiated carbonates in the dataset; as such, carbonates were not further subdivided for analysis.

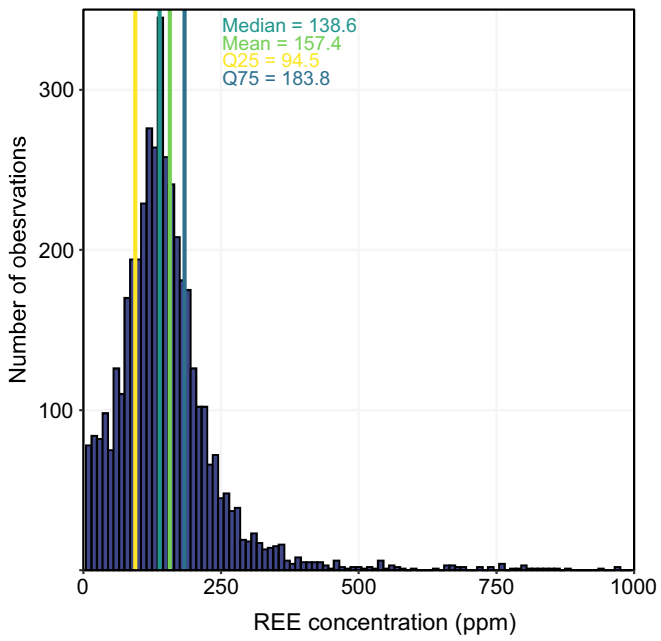


Fig. 1. Histogram displaying the distribution of total REE concentrations for samples used in this study with median, mean, and first and third quartiles indicated. Since only 15 points are above 1000 ppm, the x-axis was trimmed to better illustrate the distribution of REE concentrations below 1000 ppm.

The concentrations of REE by lithology are displayed in Fig. 2A. Kruskal-Wallis statistical testing indicated these differences are statistically significant (p-value <0.05). Fig. 2B highlights the variation in REE abundances through time in this dataset and Kruskal-Wallis testing also indicated statistically significant differences between geologic periods.

3.1.1. Correlation analysis

The correlograms for both the untransformed and CLR transformed data (Fig. 3A and B) indicate that REE were most strongly correlated with Th, while they were weakly correlated (>2050%) with Al, Cs, Fe, Ga, Hf, K, Na, Nb, P, Rb, Ti, and Zr in the untransformed data and Hf, Nb, P, and Zr in the CLR transformed data. Conversely, REE were most strongly anti-correlated with Ca and Mg, common constituents of carbonate rocks, which is consistent with Fig. 2 that shows carbonates typically contain significantly lower REE abundances. Additionally, positive correlations were found among lithophile elements (e.g., Al, Cs, Ga, Hf, K, Na, Nb, REE, Th, Ti, and Zr), elements mobilized by oxidative or sulfide weathering that subsequently accumulate in fine-grained organic-rich sediments (e.g. Ni, U, and V), and elements commonly associated with carbonate rocks (Ca, Mg, and Sr). Although the relative magnitude of the correlations differs between the untransformed and transformed data, both showed similar relationships between variables.

3.1.2. Principal component analysis

Principal component analysis performed on the CLR-transformed data revealed that 12 principal components (PCs) accounted for >95% of the variance in the data, while 76% of the variance was explained by the first four PCs (Fig. 4A).

Table 1
Summary statistics for total REE concentration data (ppm) by lithology for Dataset 1. 61 samples did not include lithological information. SD = standard deviation. MAD = mean absolute deviation.

Lithology	n	Min	Q1	Median	Mean	Q3	Max	SD	MAD
All data	4364	1.4	94.5	138.6	157.4	183.8	8170	225.7	66.0
Carbonate	490	1.5	22.4	56.1	75.6	103.4	1913.0	113.3	55.3
Chemical	19	2.1	2.7	28.0	43.7	67.4	175.9	52.3	38.0
Clastic	3760	1.4	109.9	144.7	168.0	189.3	8170.0	232.9	58.4
Phosphorite	34	13.9	227.5	375.9	474.9	725.1	1443.2	320.3	289.0
Unclassified	61	–	–	–	–	–	–	–	–
Carbonate (undifferentiated)	118	5.9	35.5	65.8	66.8	88.4	298.6	42.9	40.1
Chert	7	28.0	34.5	64.6	69.5	74.3	175.9	51.0	43.5
Conglomerate	12	12.9	23.2	41.8	75.4	94.5	254.6	77.9	31.5
Dolomite	45	1.7	5.7	22.3	46.3	60.1	341.2	66.1	27.1
Evaporite	12	2.1	2.4	3.2	28.7	24.3	138.1	48.8	1.3
Limestone	327	1.5	21.0	57.0	82.8	118.7	1913.0	133.5	64.0
Mudstone	945	1.4	116.6	162.3	184.5	216.8	2143.9	141.4	74.1
Sandstone	209	18.7	50.6	101.5	115.7	140.5	1566.1	130.4	66.7
Shale	2308	2.7	110.5	142.1	165.6	181.0	8170.0	277.3	51.6
Siltstone	286	9.1	130.6	155.6	175.4	203.8	1169.3	105.5	48.6

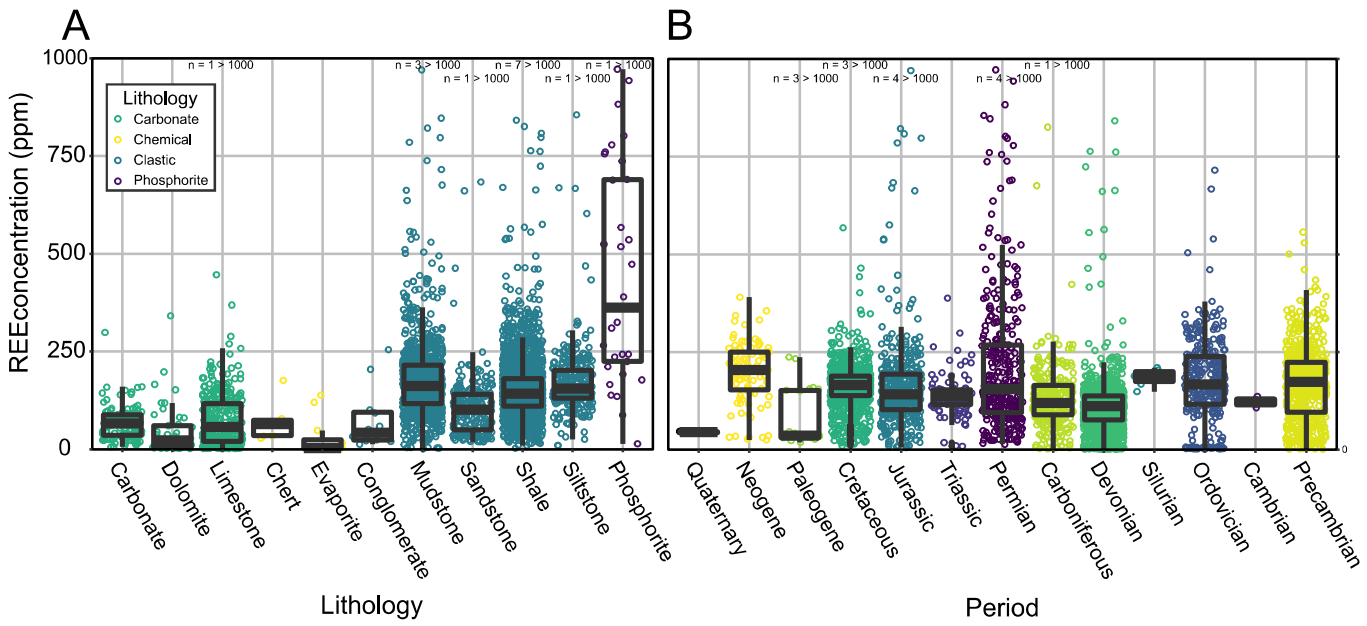


Fig. 2. A - Boxplot of total REE concentration by lithology for data used in this study. B – Boxplot of total REE concentration through time for data used in this study. The box represents the first and third quartiles, while the centre line represents the median. The whiskers extend to 1.5 times the interquartile range, with points beyond the whiskers being considered outliers. The number of observations >1000 ppm is indicated above each box.

The PCA biplot of PC1 vs PC2 (Fig. 4B) indicated the variables were essentially divided into three groups, the first of which comprised Ca, Mg, and Sr: elements that contribute to PC1 (Fig. 4C) and are common constituents of carbonate rocks. The second group, representative of PC2, included Ba, Ni, U, and V (those related to oxidative weathering), while REE and the remaining variables, including the incompatible elements, plotted in the upper-right quadrant and had some contribution to each of the PCs. The elements in each of these groups were also highly correlated among one another in the correlation analysis, showing agreement between both methods in identifying elemental associations. The confidence ellipses grouped the samples based on their lithologies, supporting the interpretation that PC1 is strongly associated with carbonates. However, this also indicated that the PCA is strongly influenced by the composition of the dataset including different lithologies. Therefore, to identify differences in principle components between the two dominant lithologies, additional PCAs were performed on clastic and carbonate samples separately (Fig. 5A-D).

Several similarities were noted between the PCA of the complete dataset and the PCA of the clastic subset, namely the three groups of variables, with an increased collinearity of P with Ca and Co with Ba in the clastic subset. The contribution of variables to each PC were also similar, with the exception of PC2 which had increased contributions from Ca, Mg, and P, while PC4 had a higher contribution from P. More considerable differences were observed in the carbonate biplot which indicated a higher collinearity of REE with Ba, Si, and V, and lower collinearity with commonly associated incompatible elements (e.g., Hf, Nb, and Th). Phosphorus showed an increased collinearity with U and Ni while Co was no longer on the same axis as Ba, Ni, U, and V. The variable contributions to each PC were similar between the clastic and carbonate subsets, apart from PC3, where P was the most important variable in the carbonate subset. The similarities in the PCA for the complete dataset compared to the clastic subset could be a result of a high proportion of the total data being labelled as clastic. Nevertheless, the results indicated differences in geochemical relationships between REE and other trace metals in clastic and carbonate rocks.

3.1.3. *Cluster analysis*

Hierarchical cluster dendrograms (Fig. 6) showed similar elemental

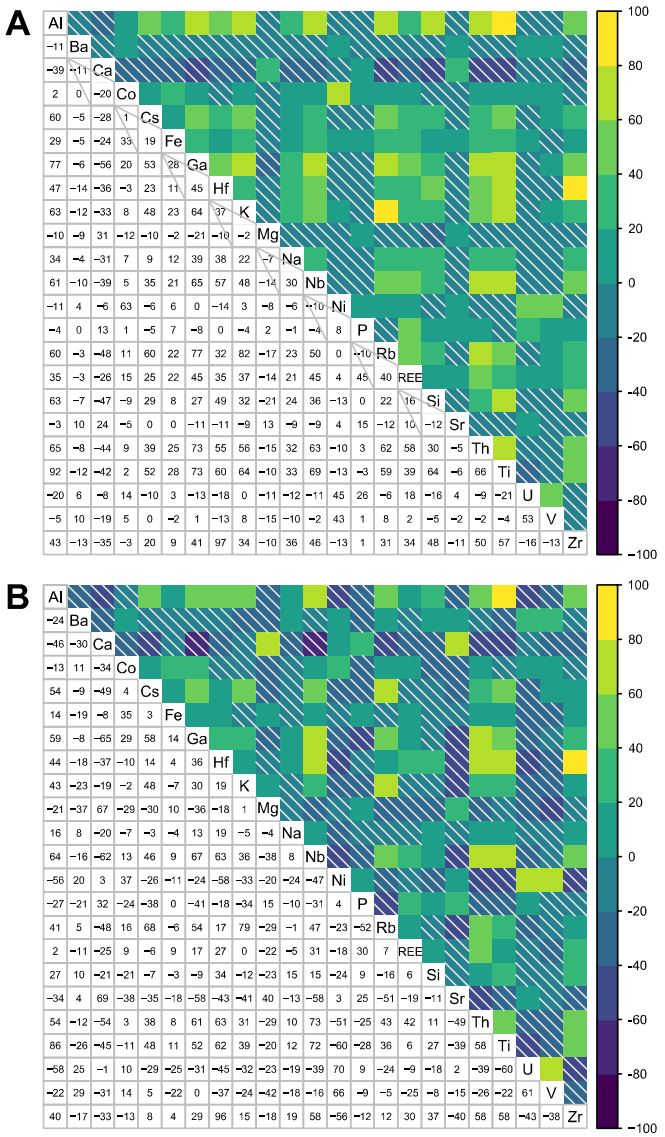


Fig. 3. Correlograms displaying the correlation coefficient for each element pair. A-untransformed data. B-CLR transformed data.

associations as revealed by the PCA. For the complete dataset and clastic subset, Ba, Ni, U, and V and Hf, P, Na, Si, and Zr had similar associations while Ca was independent of any group. In both cases, REE were most closely associated with Th and Nb. Nickel, P, U, and V, and Ca and Sr were associated in both the carbonate PCA and cluster dendrogram. However, there were some notable differences between the two algorithms, as the dendrogram suggested REE-Fe and Ba—Mg associations.

3.2. Supervised machine learning

Seven machine learning algorithms were implemented to predict REE concentrations under several scenarios and their relative accuracies compared to each other were assessed. The algorithms, the details for each scenario, and results are summarized in Table 2. Model parameters are available in Table SI3. Scenario A was performed on the complete dataset. Scenario B only considered predictions using major elements (Al, Ca, Fe, K, Mg, Na, P, Si, Ti, and Zr) to assess the accuracy of the MLAs as some historical datasets may not include trace metal data. Scenario C considered the top five elements which were most correlated to REE (Ga, Nb, P, Rb, and Th) as increasing the number of predictor variables can affect the computing power required to run the models. As demonstrated by the statistical analysis and unsupervised learning, there can be significant differences in the REE concentration between clastic and carbonate lithologies, therefore the MLAs were performed by separating the dataset into clastic and carbonate lithologies (Scenario D and E, respectively). Finally, only data from Alberta, Canada was incorporated to assess whether the machine learning models are more accurate for an individual region, a possible reflection of the importance of local geologic history (Scenario F).

Based on the RMSE values, typically the most accurate models were GB, AB, and RF, whereas the SVM was the least accurate model for all scenarios. Both regression models, ridge and linear, performed well for Scenario B. GB performed better than AB for all scenarios, which was anticipated since it is an improved boosting model. All models had the highest accuracy when considering only the carbonate subset (Scenario E). The performance of five models, KNN, LR, RR, RF, and SVM were the poorest for the Alberta dataset (Scenario F), while the boosting algorithms had the lowest performance for Scenario B, which included only major elements as predictors. With the exception of KNN and SVM, the models were more accurate when more variables were involved (e.g., Scenario A vs Scenario B or C) implying there is a trade-off between decreasing the number of variables and overall model accuracy. However, only a minor decrease in accuracy was observed between Scenario A and Scenario C indicating that selecting only a few related variables can decrease computational power while having a limited effect on model performance. The most important variable for predicting REE concentration for AB, GB, LR, RR, and three of the RF models was P, while Th was important for the remaining three RF models as well as two SVM and one RR model. Additionally, Ca, Fe, Hf, Nb, Rb, Si, and Zr were also shown to be important predictors across several MLAs. The regression models and KNN were the computationally fastest models to run, followed by SVM, RF, and the boosting models, indicating a compromise between computation time and model accuracy.

4. Discussion

4.1. Machine learning

4.1.1. Identifying geochemical relationships using unsupervised machine learning

Overall, the correlation, PCA, and cluster analyses indicated similar geochemical relationships. Elements typically associated with carbonate rocks (i.e., Ca, Mg, and Sr) were highly correlated and showed close relationships in the PCA and cluster analysis, as were elements that are typically incompatible in magmatic systems. For the complete dataset and clastic subset, REE were most closely associated with incompatible elements, specifically Th and Nb. Conversely, REE were associated with Al, Fe, and Ti in the carbonate cluster analysis. In a study of metal distribution patterns in the Campania region of Italy, it was observed that REE had the highest correlation with Th and were negatively correlated with Mg (Ambrosino et al., 2022). Hence, the authors proposed that the main source of REE was a result of the presence of Fe—Mn oxides and hydroxides associated with clay minerals (Ambrosino et al., 2022), which is supported by the association with Al and other lithophile elements shown in the unsupervised machine learning. In a study of soil from Turkey, LREE and HREE were also found to be correlated with incompatible elements, specifically Hf, Th, and Zr, while a PCA of the data showed the first PC was composed of As, Ba, Hf, Mo, Nb, Pb, REE, Th, and Zr indicating these elements may have a related origin or enrichment process (Vural, 2020). Similarly, Bishop et al. (2023) investigated metal associations in a global compilation of coal combustion by-products using correlation analysis, cluster analysis, and PCA and found that Nd, a proxy for REE, was most associated with Al, Th, Ti, and Zr. Overall, these wide-ranging studies indicate that REE typically have similar geochemical associations in sedimentary environments, specifically with Th, and that enrichment of associated elements could be indicative of elevated REE concentrations. Since REE are typically

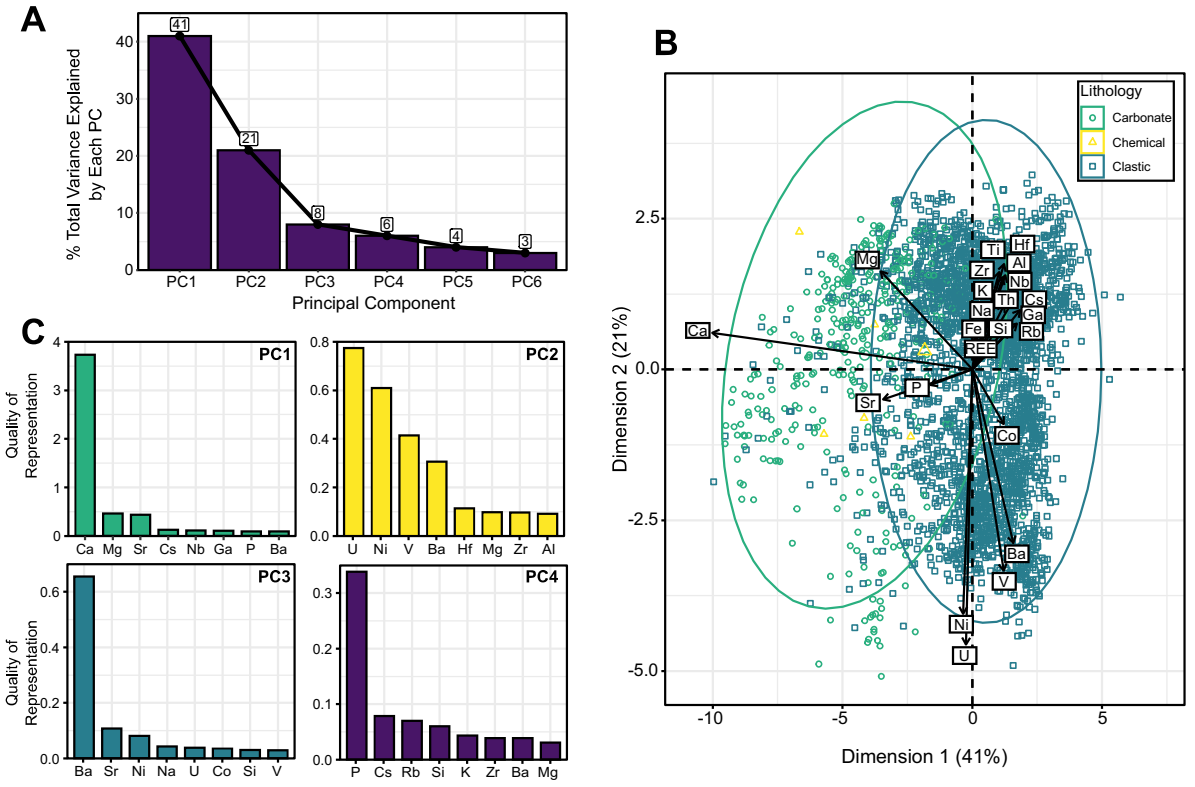


Fig. 4. A – Screeplot displaying each PC's contribution to variance within the dataset. B – PCA biplot of Dimension 1 vs 2 for the complete dataset including 95% confidence ellipses for carbonate and clastic lithologies. C – Variable contribution to each of the first four PCs.

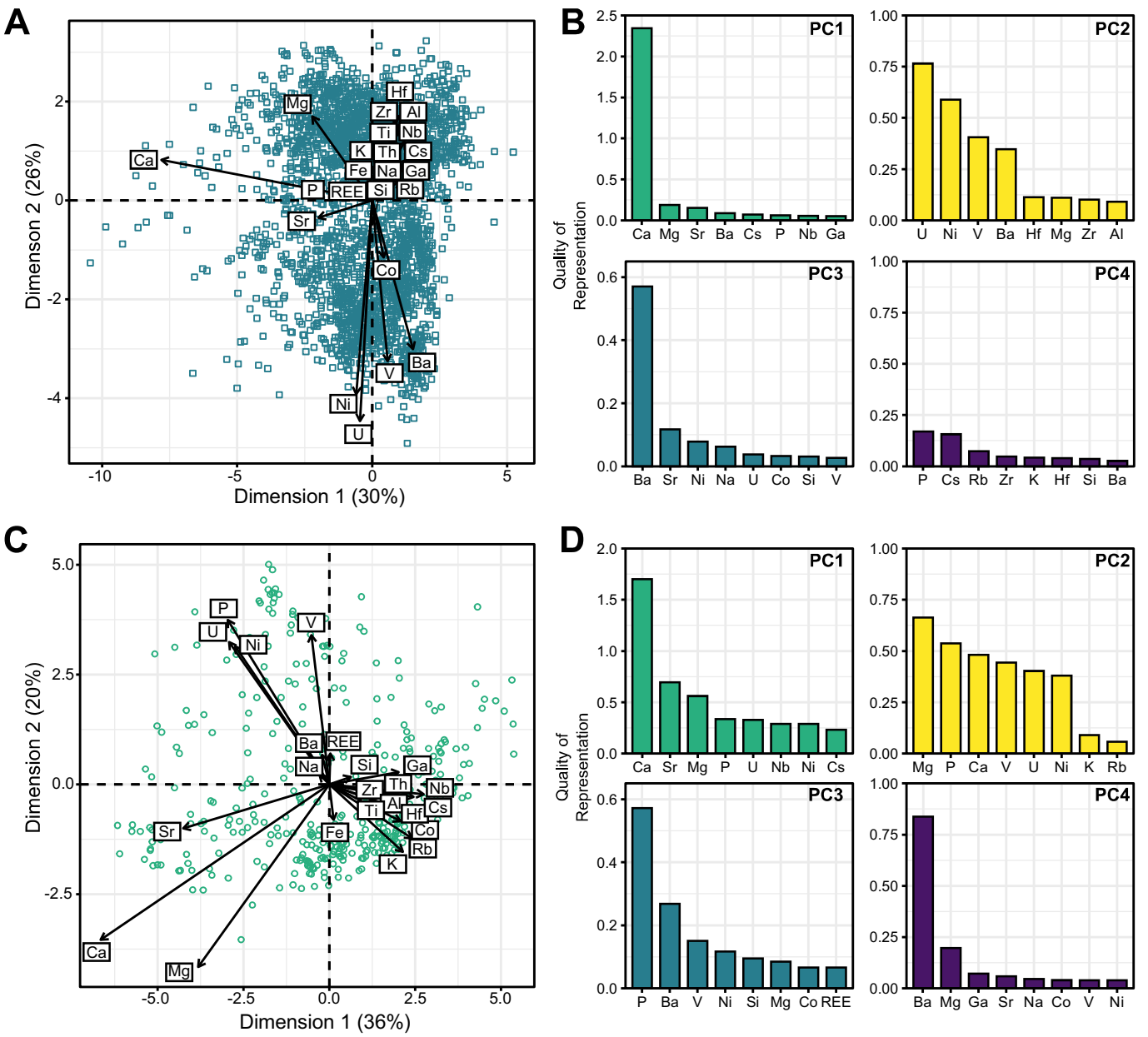


Fig. 5. A - PCA biplot for clastic labelled samples. B – Variable contribution to each PC for the clastic labelled samples. C - PCA biplot for carbonate labelled samples. D – Variable contribution to each PC for the carbonate labelled samples.

associated with incompatible elements in geologic environments, this work further demonstrates the geologic relevance of the unsupervised machine learning used to identify elemental relationships with REE.

Although these models assume the data was representative of the total population, that was not necessarily the case here. Similarities between the complete dataset and clastic subset, and disparities between the complete dataset and carbonate subset, are likely a result of bias, since ~86% of samples were labelled as clastic. Accordingly, subsequent data analysis should be performed on samples from broadly similar geologic environments, as was done by separating the clastic and carbonate lithologies. Similar findings using correlation and cluster analysis for coal geochemistry indicated that samples should originate from the same horizon since geological factors have the potential to influence the relationship between elements (Eskanazy et al., 2010).

4.1.2. Predicting REE abundances and comparison of supervised machine learning models

The supervised machine learning results indicated that RF and GB performed comparatively well and were the most accurate models based on RSME values in five of the six scenarios considered. The relatively higher performance of RF and boosting algorithms in geological contexts has been previously observed. For instance, Buccione et al. (2023) used ANN, SVM, RF, and XGBoost to predict HREE distributions in southern Italian karst bauxite deposits based on major oxide abundances and found that the XGBoost and RF algorithms had the highest accuracy. For predicting lithology type based on well logs, ensemble methods including RF and boosting had better performance than SVM, Naïve Bayes, and ANN (Xie et al., 2018). Similarly, studies using supervised machine learning to predict land cover type and mineral prospectivity have demonstrated that RF algorithms can be more accurate and require less computing power than SVM and ANN algorithms and can overcome the "black-box" limitations of ANN (Rodriguez-Galiano et al., 2015;

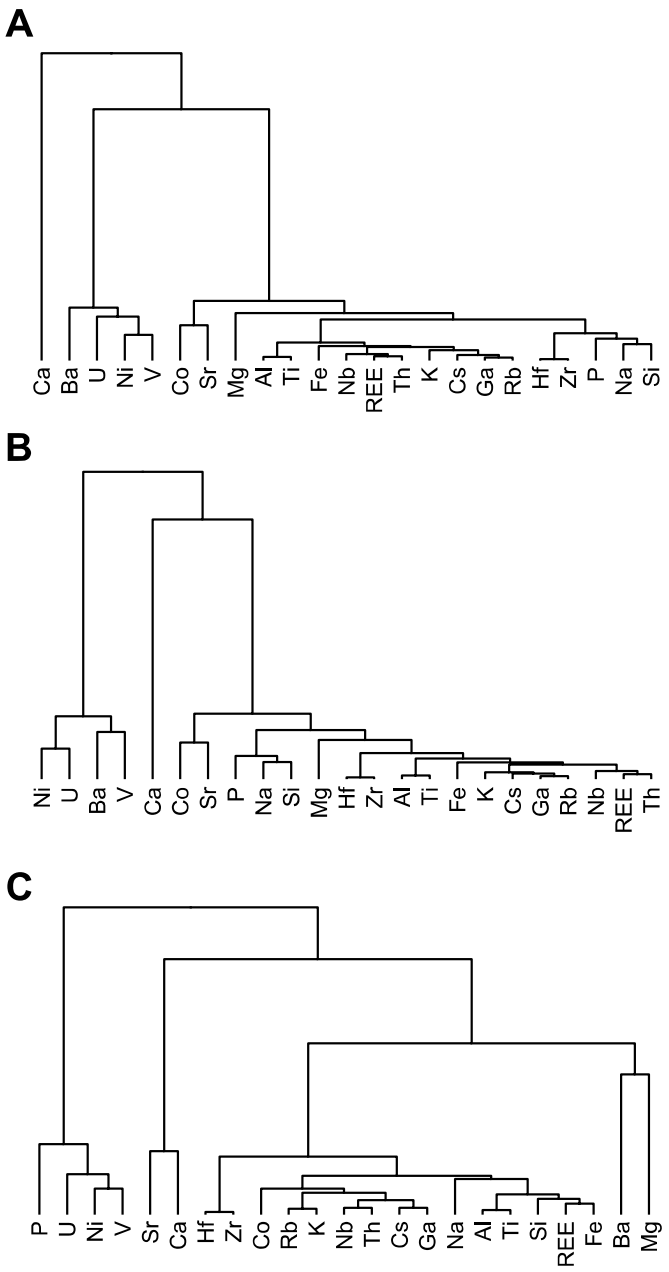


Fig. 6. Hierarchical cluster dendrograms for A: all data, B: clastic subset, and C: carbonate subset.

Rodriguez-Galiano and Chica-Rivas, 2014). For mapping mineral prospectivity at the continental scale for MVT and CD deposits, Lawley et al. (2022) found GB to have the best performance.

Typically, the most important features used to predict REE concentrations were P and Th, regardless of lithology, and these were both found to be highly associated with REE in the unsupervised machine learning. This could reflect REE incorporation into P-bearing phases such as calcium fluorapatite, monazite, or xenotime, or with incompatible and detrital mineral phases. However, there could be other elements that are important REE predictors but were not included in the dataset considered here, such as U. Additionally, MLAs were not optimized for each model which could have an impact on the overall accuracy of each algorithm, but instead provide a starting point for identifying the most suitable MLA for predicting elemental concentrations.

All models had the best performance for the carbonate subset (Scenario E) compared to the complete dataset (Scenario A). This is similar to the findings of Lawley et al. (2022) where geochemical models for Clastic Dominated (CD) and Mississippi Valley Type (MVT) deposits trained on samples from Canada and the USA did not generalize to Australia, which could be the result of different geologic processes that are separated by space and time. Here, the accuracy for the Alberta subset was generally low, which may stem from the inclusion of both carbonate and clastic lithologies and the geologic history of the Western Canada Sedimentary Basin, which spans from the Precambrian to the

Table 2
Machine learning results for predicting REE concentrations for each of the seven models. RMSE = root mean squared error. FI = feature importance; the top three variables which were most important for predicting REE concentrations.

Scenario and description	Model	RMSE	FI 1	FI 2	FI 3
A - Complete dataset	Adaboost	67.7	P	Th	Nb
	Gradient Boosting	62.3	P	Th	Nb
	Linear Regression	68.7	P	Th	Zr
	Ridge Regression	68.7	P	Th	Zr
	KNN	87.0	Si	Al	Ca
	Random Forest	65.3	Th	P	Sr
	Support Vector Machine	99.3	Th	Rb	Ga
B - Major elements (Al, Ca, Fe, K, Mg, Na, P, Si, Ti, Zr)	Adaboost	82.2	P	Ca	Fe
	Gradient Boosting	78.2	P	Ca	Fe
	Linear Regression	76.3	P	Si	Ca
	Random Forest	74.5	P	Zr	K
	Ridge Regression	76.3	P	Si	Ca
	Support Vector Machine	100.7	P	K	Zr
	KNN	87.1	Si	Al	Ca
C - Top 5 correlated elements (Ga, Nb, P Rb, Th)	Adaboost	69.4	P	Th	Nb
	Gradient Boosting	65.2	P	Th	Nb
	KNN	81.5	P	Rb	Nb
	Linear Regression	70.3	P	Th	Nb
	Random Forest	66.2	P	Th	Nb
	Ridge Regression	70.3	P	Th	Nb
	Support Vector Machine	97.6	Rb	Ga	Th
D - Clastic subset	Adaboost	77.8	P	Nb	Th
	Gradient Boosting	66.9	P	Th	Sr
	Linear Regression	69.4	P	Th	Zr
	Ridge Regression	69.4	P	Th	Zr
	KNN	88.4	Si	Al	Ca
	Random Forest	64.2	Th	P	Nb
	Support Vector Machine	99.1	Th	Rb	Nb
E - Carbonate subset	KNN	63.4	Ca	Si	Al
	Adaboost	64.9	Fe	P	Nb
	Linear Regression	56.6	Hf	Zr	Th
	Gradient Boosting	48.9	P	Fe	Th
	Random Forest	50.3	P	Th	Zr
	Support Vector Machine	70.6	P	Th	Ga
	Ridge Regression	56.4	Th	U	Fe
F - Alberta subset	KNN	108.9	Ca	Si	Al
	Linear Regression	85.2	Hf	Zr	Th
	Support Vector Machine	123.9	Nb	Th	Rb
	Adaboost	82.1	P	Sr	Th
	Gradient Boosting	77.2	P	Th	Sr
	Random Forest	84.1	Th	P	Sr
	Ridge Regression	85.2	Zr	Hf	Th

Present. Accordingly, having increased similarity between factors such as lithology, locality, and geologic setting could be important for increasing predictive model accuracy and identifying meaningful relationships using unsupervised MLAs.

4.2. Rare earth element abundances and relationships in sedimentary strata

Occurrences of REE in sedimentary systems can either be primary, which involves concurrent or *syn*-depositional emplacement of REE-rich material, or secondary, which involves post-depositional processes (Creason et al., 2023). The data analysis indicated that phosphorites and clastic lithologies were significantly more concentrated in REE, which can occur through both primary and secondary processes, than carbonate and evaporite lithologies, which are typically enriched through primary processes. Phosphorites have been shown to concentrate REE during their formation, primarily as a result of diagenetic processes where REE substitutes for Ca in francolite (Emsbo et al., 2015; McArthur and Walsh, 1984). McArthur and Walsh (1984) postulated that the REE are sourced from pore water through diagenetic reactions of occluded and enclosing sediment, and the majority of REE is incorporated into the phosphorites post-depositionally with older samples having higher REE concentrations since they have more time to incorporate metals. Conversely, Emsbo et al. (2015) favoured a model for REE incorporation from seawater with secular variations in seawater chemistry through time leading to differences in the REE content of phosphorites and no evidence for increasing REE concentration as a function of time.

Fine-grained clastic lithologies, including mudstones and shales, were also observed to contain elevated REE concentrations. The main contributor of REE to marine sediments and the ocean is particulate scavenging of adsorbed REE which may be transported nearly in bulk from source to sediment (Condie, 1991). Weathering and erosion result in very limited concentrations of REE in solution, as only a few percent of REE entering the oceans are in the dissolved phase, while the bulk of REE are transported as eroded particulate phases, likely clay minerals (Fleet, 1984). A substantial proportion of the REE in clays are loosely held and are available to take part in exchange reactions, and the dominant control on the REE concentrations in seawater is adsorptive scavenging of REE by particles (Elderfield and Greaves, 1982). Conversely, low REE concentrations in eroded material with quartz and other major silicates are reflected in the lower REE contents of sandstones relative to more clay-rich lithologies such as shales (Fleet, 1984). Although there has been some debate surrounding the main source of REE to the ocean: either iron-oxyhydroxide or clay minerals, recent research has asserted that the latter is responsible for the majority of the REE flux to the oceans (Abbott et al., 2019), which is not unexpected given the high affinity that clay minerals have for REE adsorption (Alshameri et al., 2019; Bradbury and Baeyens, 2005). Although sandstones typically have lower concentrations of REE, in some instances the incorporation of REE-enriched minerals including zircon or monazite could increase their overall concentration.

Carbonate lithologies were found to have, on average, the lowest concentration of REE among those studied here. Aluminum, Ga, Hf, REE, and Th are not incorporated into the carbonate lattice, and are therefore an ideal proxy for terrigenous material, whereas elements such as Cr, Cu, Mg, Mo, Na, Ni, P, Sr, V, U, and Zn can be incorporated into carbonates and mirror the chemical composition of paleoseawater (Mirza et al., 2021). In part, this reflects the authigenic processes through which carbonates precipitate from seawater, either biotically or abiotically. Given the generally low concentrations of REE in seawater, and the strong influence of particulate scavenging, it is therefore expected that carbonates would generally have low REE abundances relative to more clastic or detritally influenced lithologies. However, several studies have noted that seawater-like REE patterns may be observed in various carbonates, and that under the right conditions, they may be used to extract information regarding paleomarine conditions (Nothdurft et al., 2004; Kamber et al., 2014; with Johannesson et al. (2006) providing an alternative view). The general lack of terrigenous input to carbonates, and their precipitation as authigenic sediments, could account for the close relationship between variables in the carbonate PCA PC1 (Ca, Mg, and Sr) and PC2 (Ni, P, U, and V), and the strong correlations between REE with Al, Ga, Hf, and Th. Overall, the relationships revealed in the PCA indicate a strong role for clay minerals and detrital input in sourcing REE to clastic deposits, and to a certain extent the influence that a lack of terrigenous input has on limiting REE concentration in authigenic carbonates.

4.3. Implications for REE prospectivity in sedimentary environments

There are several secondary sources derived from sedimentary environments that have the potential to host elevated REE abundances which could be exploited as a potential resource, including: (i) coal and coal combustion by-products (CCBs); (ii) phosphorites and phosphogypsum; (iii) deep-sea muds; (iv) oil sands tailings; and (v) geothermal and formation waters. Due to a combination of elevated REE concentrations coupled with a high volume of available source materials globally, these secondary streams may have the potential to meet current and future global demand (Gaustad et al., 2021). Coal could be a source of REE since the concentrations in some deposits can be equal to or higher than those from conventional ores, especially those with a high volcanic ash content (Seredin and Dai, 2012). However, CCBs are emerging as a more likely source since REE can be significantly concentrated in the ashes, there are strong incentives for reuse since CCBs are an environment liability, they are readily available around the world, particle sizes are small which reduces the need for crushing and grinding, and radioactive tailings are significantly reduced (Blissett et al., 2014; Taggart et al., 2016; Fu et al., 2022). Resource evaluation of ash from Powder River Basin coals in the USA demonstrated that those with a favourable geochemistry for extraction could be economically promising and a near term economic resource (Bagdonas et al., 2022). CCBs in numerous countries including Brazil (Lange et al., 2017), Canada (Bishop et al., 2022, 2023), China (e.g., Dai et al., 2014; Wang et al., 2019; Zhang et al., 2019; Wu et al., 2022; Hu et al., 2023), India (Modi et al., 2021; Sandeep et al., 2023), Poland (Blissett et al., 2014; Franus et al., 2015), South Africa (Wagner and Matiane, 2018), the United Kingdom (Blissett et al., 2014), and the USA (e.g., Taggart et al., 2016; Kolker et al., 2017; Huang et al., 2020; Mastalerz et al., 2022) have been investigated for their REE recovery potential. This has spurred the development of extraction processes, although most are still in the proof of concept and pilot phases (Dodbiba and Fujita, 2023).

Phosphorites in the United States could also provide a significant supply of REE, with the value of the REE sometimes outstripping the value of the produced P (Emsbo et al., 2015). High recovery rates which apply the same processes used to produce phosphate further increases the suitability of phosphorites for REE recovery especially since they can be produced as a by-product (Emsbo et al., 2015) and, importantly, unlike carbonatite deposits, they are enriched in HREE which are less abundant than LREE (Hein et al., 2016). Additionally, phosphogypsum, a waste product from fertilizer production, could also be a promising source of REE due to large volumes and the relative ease of recovery (Cánovas et al., 2019). As such, understanding enrichment processes and predicting REE concentrations may prove useful for the development of phosphorite or phosphogypsum-based extraction strategies.

Deep-sea muds can contain comparable REE concentrations to ion adsorption clay deposits of China, elevated HREE abundances, low Th and U concentrations, and can be readily recovered through leaching with dilute acids (Kato et al., 2011; Takaya et al., 2018). However, deep-sea mining is anticipated to create environmental impacts through: the removal of the resource (e.g., nodules and crusts) which can host unique fauna, geochemical and physical changes to the seafloor, sediment plumes, contaminant release, and increase in sound, vibration, and light which can affect biodiversity and potentially lead to extinction of rare

species (Levin et al., 2020). While this may not be a REE recovery strategy developed in the near term, generating knowledge of REE distributions and the ability to predict enrichments is an important first step in assessing future feasibility.

By-products from hydrocarbon production have been studied as a novel REE source as they can host elevated metal concentrations. High REE concentrations have been found in the fine tailings from oil sands operations, with these tailings being considered an environmental liability so REE recovery could be included as part of remediation efforts (Roth et al., 2017). Formation waters and basinal brines, co-produced with hydrocarbons or geothermal energy have also been subject to investigation for REE since their recovery can offset operating expenses (Smith et al., 2017; Miranda et al., 2022). However, since these waters are generally brines with high TDS values and low REE concentrations, few studies have been able to accurately quantify REE in these fluids (Kokh et al., 2021). The speciation and abundance of REE in formation waters is strongly dependent on temperature, pH, ionic strength, and the presence of complexing anions (Lewis et al., 1998), with pH being the most important geochemical factor as more acidic geothermal and formation fluids contain significantly higher dissolved REE concentrations (Liu et al., 2016; Quillinan et al., 2018). However, Quillinan et al. (2018) postulated that geologic controls and basin history may have the most influence on REE content. Water with low REE concentrations that infiltrates into formations containing high abundances of REE may equilibrate with the surrounding environment, potentially releasing bound REE into solution through water-rock interactions (Möller et al., 2021). Therefore, waters percolating through REE-rich sediments may have increased REE concentrations relative to waters percolating through REE-poor sediments. Accordingly, by analyzing available rock chemistry data, intervals which have the potential to host brines with the highest REE concentrations can be predicted. This has been demonstrated for the Western Canada Sedimentary Basin, where core samples from the Jurassic and Cretaceous periods contained the highest REE concentrations (Fig. 7). Correspondingly, the highest REE concentrations measured during a brine sampling study in the Saskatchewan portion of the basin were observed in the Upper Cretaceous Belly River Formation and the Jurassic Shaunavon Formation, both of which have high proportions of fine-grained sediments (Bishop et al., 2024). Since there is significantly more REE data for rock samples than formation waters, identifying enriched lithologies could be the first step in guiding future sampling programs.

Elucidating the geochemical and lithological controls on REE abundances in sedimentary basins is important to determine geochemical indicators of REE enrichment, and by extension, the mineral potential. Understanding the geochemical and lithological factors that affect REE concentrations in sedimentary successions is similarly important for furthering our understanding of behaviour of REE in sedimentary environments and the implications for variations over geological timescales. The data analysis presented here further confirms that phosphorites contain the highest REE concentration relative to other lithologies. Similarly, elemental relationships observed between unsupervised machine learning here and that from CCBs in Bishop et al. (2023) indicate that this work, which investigates REE from a broad range of sedimentary environments, can be applied to a narrower range of applications such as coal and CCB studies.

Identifying REE enriched lithologies can assist in finding intervals which may host sediments or fluids with elevated concentrations. These elemental relationships could be important for assessing a given environment for REE enrichment, and supervised machine learning could be utilized to estimate the concentration based on more readily available major and trace element data. This could be crucial for use with historical datasets that may not include REE and could reduce costs associated with performing preliminary geochemical analyses. The methodologies detailed here could be used to target stratigraphic intervals based on historical datasets, limiting the need for broad, untargeted sampling. Although REE concentrations are typically much lower than what is found in ore deposits, the volumes of available feedstock, in addition to potentially less expensive and environmentally damaging processes, could make secondary REE sources attractive for meeting future demand of these critical metals. Such an approach to REE extraction could play a significant role toward the establishment of a circular economy, and these proposed secondary sources are more globally distributed than traditional REE deposits which could prove to be important domestic sources. However, the potential for each of these to become a resource will depend on the development of an effective extraction technology, and ultimately economic conditions including demand and commodity price.

5. Conclusions

Increasing the supply of critical metals necessary in clean energy technology, including REE, is imperative as society strives to decarbonize the economy. As demand increases and major metal deposits become depleted, new tools, such as machine learning, and supplies,

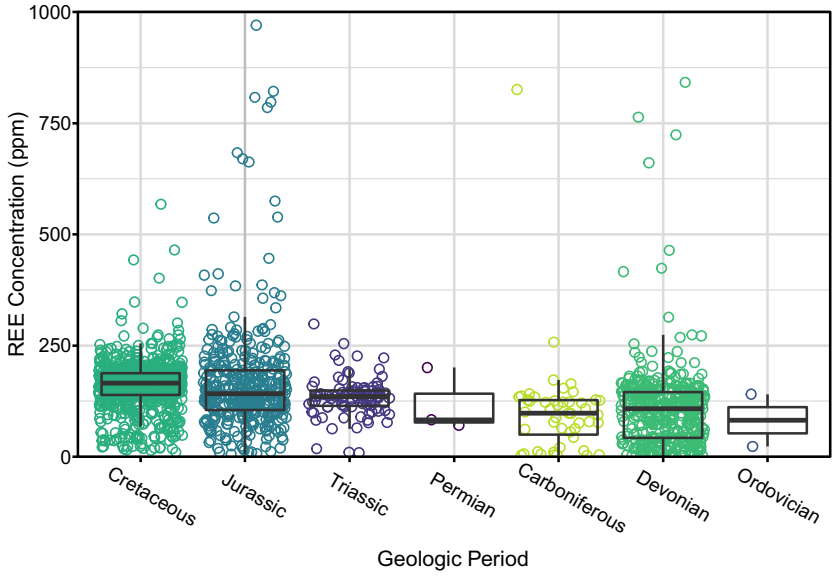


Fig. 7. Distribution of REE concentrations by period for sedimentary strata in the Western Canada Sedimentary Basin.

such as those from secondary sources, could play important roles in meeting this need. This work explored secondary sources of REE from sedimentary environments by applying compositional data analysis principles in tandem with unsupervised MLAs to discern geochemical indicators of REE enrichments. It was found that REE were most enriched in phosphorites and fine-grained clastic lithologies, while the three unsupervised models showed agreement between each other, demonstrating that REE were most associated with incompatible elements (i.e. Nb, Th, and Zr) and P. Additionally, supervised MLAs were utilized to predict REE concentrations under several scenarios, with AB, GB, and RF models having the highest performance. The most important features for predicting REE concentrations were Th and P. The associations between REE and Th and P found in both types of MLAs are geologically relevant since REE have been found to be associated with Th in fine-grained lithologies and they are commonly incorporated into phosphate minerals (monazite and xenotime). However, geologic controls can influence these MLAs and therefore, datasets should include samples from similar geologic environments to provide the most geologically relevant results and highest model performance.

This work additionally contributes to the understanding of the controls and distribution on REE in sedimentary strata. Findings from this work can be incorporated into broader frameworks that may include additional geochemical, spatial, historical, drill core, or other data as part of an exploration strategy into secondary environments to identify targets with the highest REE potential, whether it be CCBs sourced from an enriched coal horizon, clay-rich tailings, or formation or geothermal waters percolating through rocks known to be high in REE. Extracting REE from secondary sources of sedimentary origin can be advantageous since there can be significant volumes of existing data, hence reducing exploration costs, and it can be part of a remediation strategy if the feedstock is a waste product – thereby contributing to the circular economy. While these sources can be promising, significant characterization is still required, and importantly, cost-effective extraction technologies must be developed in order to make these secondary sources economically viable. Meeting the metal needs of the energy transition is crucial for limiting the effects of climate change, with secondary sources playing a potentially critical role.

CRediT authorship contribution statement

Brendan A. Bishop: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Leslie J. Robbins:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare no conflict of interest in the completion of this work.

Data availability

Data will be made available on request.

Acknowledgements

The authors acknowledge the insights and thoughtful discussion of Daniel Gregory (University of Toronto) and Kaarel Mänd (University of Tartu) and comments from an anonymous reviewer that helped to improve this work. Financial support was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) in the form of a Canada Graduate Doctoral Scholarship (BAB: CGS—D) and a Discovery Grant (LJR: RGPIN-2021-02523).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gexplo.2024.107388.

References

Abbott, A.N., Löhr, S., Trethewy, M., 2019. Are clay minerals the primary control on the oceanic rare earth element budget? Front. Mar. Sci. 6, 504. https://doi.org/10.3389/fmars.2019.00504.

Ahmed, A.D., Hood, S.B., Cooke, D.R., Belousov, I., 2020. Unsupervised clustering of LA-ICP-MS raster map data for geological interpretation: a case study using epidote from the Yerington district. Nevada. Apple. Comput. Geosci. 8, 100036 https://doi.org/10.1016/j.acags.2020.100036.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London, p. 416.

Alshameri, A., He, H., Xin, C., Zhu, J., Xinghu, W., Zhu, R., Wang, H., 2019. Understanding the role of natural clay minerals as effective adsorbents and alternative source of rare earth elements: adsorption operative parameters. Hydrometallurgy 185, 149–161. https://doi.org/10.1016/j.hydromet.2019.02.016.

Ambrosino, M., Albanese, S., De Vivo, B., Guagliardi, I., Guarino, A., Lima, A., Cicchella, D., 2022. Identification of rare Earth elements (REEs) distribution patterns in the soils of Campania region (Italy) using compositional and multivariate data analysis. J. Geochem. Explor. 243, 107112 https://doi.org/10.1016/j.gexplo.2022.107112.

Bagdonas, D.A., Enriquez, A.J., Coddington, K.A., Finnoff, D.C., McLaughlin, J.F., Bazilian, M.D., Phillips, E.H., McLing, T.L., 2022. Rare earth element resource evaluation of coal byproducts: a case study from the Powder River Basin, Wyoming. Renew. Sust. Energ. Rev. 158, 112148 https://doi.org/10.1016/j.rser.2022.112148.

Balaram, V., 2019. Rare earth elements: a review of applications, occurrence, exploration, analysis, recycling, and environmental impact. Geosci. Front. 10, 1285–1303. https://doi.org/10.1016/j.gsf.2018.12.005.

Bern, C.R., Birdwell, J.E., Jubb, A.M., 2021. Water–rock interaction and the concentrations of major, trace, and rare earth elements in hydrocarbon-associated produced waters of the United States. Environ Sci Process Impacts 23, 1198–1219. https://doi.org/10.1039/D1EM00080B.

Bhattacharya, S., Carr, T.R., Pal, M., 2016. Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA. J. Nat. Gas Sci. Eng. 33, 1119–1133. https://doi.org/10.1016/j.jngse.2016.04.055.

Bhuiyan, M., Esmaieli, K., Ordóñez-Calderón, J.C., 2019. Application of data analytics techniques to establish geometallurgical relationships to bond work index at the Paracutu Mine, Minas Gerais, Brazil. Minerals 9. https://doi.org/10.3390/min9050302.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7.

Bishop, B.A., Jensen, G.K.S., Robbins, L.J., 2022. Rare Earth Element Abundances in Coal Combustion Byproducts of Saskatchewan. In: Summary of Investigations 2022, Volume 1, Saskatchewan Geological Survey, Saskatchewan Ministry of Energy and Resources, Miscellaneous Report 2022–4.1, Paper A-2.

Bishop, B.A., Shivakumar, K.R., Alessi, D.S., Robbins, L.J., 2023. Insights into the rare earth element potential of coal combustion by-products from western Canada. Environ. Sci. Adv. 2, 529–542. https://doi.org/10.1039/D2VA00310D.

Bishop, B.A., Jensen, G.K.S., Alessi, D.S., Robbins, L.J., 2024. Investigating the Critical Mineral Potential of Saskatchewan Formation Waters: Results from the 2021 and 2022 Sampling Programs. In: Summary of Investigations 2024, Volume 1, Saskatchewan Geological Survey, Saskatchewan Ministry of Energy and Resources, Miscellaneous Report 2024-4.1, Paper A-1.Regina, SK.

Blissett, R.S., Smalley, N., Rowson, N.A., 2014. An investigation into six coal fly ashes from the United Kingdom and Poland to evaluate rare earth element content. Fuel 119, 236–239. https://doi.org/10.1016/j.fuel.2013.11.053.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. Presented at the COLT92: 5th Annual Workshop on Computational Learning Theory. ACM, Pittsburgh Pennsylvania USA, pp. 144–152. https://doi.org/10.1145/130385.130401.

Bradbury, M.H., Baeyens, B., 2005. Modelling the sorption of Mn(II), Co(II), Ni(II), Zn (II), Cd(II), Eu(III), Am(III), Sn(IV), Th(IV), Np(V) and U(VI) on montmorillonite: Linear free energy relationships and estimates of surface binding constants for some selected heavy metals and actinides. Geochim. Cosmochim. Acta 69, 875–892. https://doi.org/10.1016/j.gca.2004.07.020.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Buccione, R., Ameur-Zaimeche, O., Ouladmansour, A., Kechiched, R., Mongelli, G., 2023. Data-centric approach for predicting critical metals distribution: Heavy rare earth elements in cretaceous Mediterranean-type karst bauxite deposits, southern Italy. Geochemistry 126026. https://doi.org/10.1016/j.chemer.2023.126026.

Cánovas, C.R., Chapron, S., Arrachart, G., Pellet-Rostaing, S., 2019. Leaching of rare earth elements (REEs) and impurities from phosphogypsum: a preliminary insight for further recovery of critical raw materials. J. Clean. Prod. 219, 225–235. https://doi.org/10.1016/j.jclepro.2019.02.104.

Caté, A., Perozzi, L., Gloaguen, E., Blouin, M., 2017. Machine learning as a tool for geologists. Lead. Edge 36, 215–219. https://doi.org/10.1190/tle36030215.1.

Chen, Y., Sui, Y., Shayilan, A., 2023. Constructing a high-performance self-training model based on support vector classifiers to detect gold mineralization-related

geochemical anomalies for gold exploration targeting. Ore Geol. Rev. 153, 105265 https://doi.org/10.1016/j.oregeorev.2022.105265.
Condie, K.C., 1991. Another look at rare earth elements in shales. Geochim. Cosmochim. Acta 55, 2527–2531. https://doi.org/10.1016/0016-7037(91)90370-K.
Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27. https://doi.org/10.1109/TIT.1967.1053964.
Creason, C.G., Justman, D., Rose, K., Montross, S., Bean, A., Mark-Moser, M., Wingo, P., Sabbatino, M., Thomas, R.B., 2023. A geo-data science method for assessing unconventional rare-earth element resources in sedimentary systems. Nat. Resour. Res. https://doi.org/10.1007/s11053-023-10163-x.
Dai, S., Zhao, L., Hower, J.C., Johnston, M.N., Song, W., Wang, P., Zhang, S., 2014. Petrology, mineralogy, and chemistry of size-fractioned fly ash from the Jungar Power Plant, Inner Mongolia, China, with emphasis on the distribution of rare earth elements. Energy Fuel 28, 1502–1514. https://doi.org/10.1021/ef402184t.
Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., Zupan, B., 2013. Orange: data mining toolbox in Python. J. Mach. Learn. Res. 14, 2349–2353.
Dodbiba, G., Fujita, T., 2023. Trends in extraction of rare earth elements from coal ashes: a review. Recycling 8, 17. https://doi.org/10.3390/recycling8010017.
Dushyantha, N., Batapola, N., Ilankoon, I.M.S.K., Rohitha, S., Premasiri, R., Abeysinghe, B., Ratnayake, N., Dissanayake, K., 2020. The story of rare earth elements (REEs): occurrences, global distribution, genesis, geology, mineralogy and global production. Ore Geol. Rev. 122, 103521 https://doi.org/10.1016/j.oregeorev.2020.103521.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300. https://doi.org/10.1023/A:1023818214614.
Elderfield, H., Greaves, M., 1982. The rare earth elements in seawater. Nature 296, 214–219.
Emsbo, P., McLaughlin, P.I., Breit, G.N., du Bray, E.A., Koenig, A.E., 2015. Rare earth elements in sedimentary phosphate deposits: solution to the global REE crisis? Gondwana Res. 27, 776–785. https://doi.org/10.1016/j.gr.2014.10.008.
Engle, M.A., Brunner, B., 2019. Considerations in the application of machine learning to aqueous geochemistry: origin of produced waters in the northern U.S. Gulf Coast Basin. Appl. Comput. Geosci. 3–4, 100012 https://doi.org/10.1016/j.acags.2019.100012.
Engle, M.A., Rowan, E.L., 2014. Geochemical evolution of produced waters from hydraulic fracturing of the Marcellus Shale, northern Appalachian Basin: a multivariate compositional data analysis approach. Int. J. Coal Geol. 126, 45–56. https://doi.org/10.1016/j.coal.2013.11.010.
Eskanazy, G., Finkelman, R.B., Chattarjee, S., 2010. Some considerations concerning the use of correlation coefficients and cluster analysis in interpreting coal geochemistry data. Int. J. Coal Geol. 83, 491–493. https://doi.org/10.1016/j.coal.2010.05.006.
European Commission, 2020. Critical Raw Materials Resilience: Charting a Path Towards Greater Security and Sustainability. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0474. Retrieved August 29, 2023.
Farrell, Ú.C., Samawi, R., Anjanappa, S., Klykov, R., Adeboye, O.O., Agic, H., Ahm, A.C., Boag, T.H., Bowyer, F., Brocks, J.J., Brunoir, T.N., Canfield, D.E., Chen, X., Cheng, M., Clarkson, M.O., Cole, D.B., Cordie, D.R., Crockford, P.W., Cui, H., Dahl, T.W., Mouro, L.D., Dewing, K., Dornbos, S.Q., Drabon, N., Dumoulin, J.A., Emmings, J.F., Endriga, C.R., Fraser, T.A., Gaines, R.R., Gaschnig, R.M., Gibson, T.M., Gilleaudeau, G.J., Gill, B.C., Goldberg, K., Guilbaud, R., Halverson, G.P., Hammarlund, E.U., Hantsoo, K.G., Henderson, M.A., Hodgskiss, M.S.W., Horner, T.J., Husson, J.M., Johnson, B., Kabanov, P., Brenhin Keller, C., Kimmig, J., Kipp, M.A., Knoll, A.H., Kreitsmann, T., Kunzmann, M., Kurzweil, F., LeRoy, M.A., Li, C., Lipp, A.G., Loydell, D.K., Lu, X., Macdonald, F.A., Magnall, J.M., Mänd, K., Mehra, A., Melchin, M.J., Miller, A.J., Mills, N.T., Mwinde, C.N., O'Connell, B., Och, L.M., Ossa Ossa, F., Pagès, A., Paiste, K., Partin, C.A., Peters, S.E., Petrov, P., Playter, T.L., Plaza-Torres, S., Porter, S.M., Poulton, S.W., Pruss, S.B., Richoz, S., Ritzer, S.R., Rooney, A.D., Sahoo, S.K., Schoepfer, S.D., Sclafani, J.A., Shen, Y., Shorttle, O., Slotznick, S.P., Smith, E.F., Spinks, S., Stockey, R.G., Strauss, J.V., Stüeken, E.E., Tecklenburg, S., Thomson, D., Tosca, N.J., Uhlein, G.J., Vizcaíno, M.N., Wang, H., White, T., Wilby, P.R., Woltz, C.R., Wood, R.A., Xiang, L., Yurchenko, I.A., Zhang, T., Planavsky, N.J., Lau, K.V., Johnston, D.T., Sperling, E.A., 2021. The sedimentary geochemistry and paleoenvironments project. Geobiology 1–12. https://doi.org/10.1111/gbi.12462.
Filzmoser, P., Hron, K., 2009. Correlation analysis for compositional data. Math. Geosci. 41, 905–919. https://doi.org/10.1007/s11004-008-9196-y.
Filzmoser, P., Hron, K., Riemann, C., 2009. Principal component analysis for compositional with outliers. Environmetrics 20, 621–632. https://doi.org/10.1007/s00180-015-0570-1.
Fix, E., Hodges, J.L., 1951. Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technical Report 4. USAF School of Aviation Medicine, Randolph Field.
Fleet, A.J., 1984. Aqueous and sedimentary geochemistry of the rare earth elements. In: Developments in Geochemistry, 2. Elsevier, pp. 343–373.
Franus, W., Wiatros-Motyka, M.M., Wdowin, M., 2015. Coal fly ash as a resource for rare earth elements. Environ. Sci. Pollut. Res. 22, 9464–9474. https://doi.org/10.1007/s11356-015-4111-9.
Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference, 1996. Presented at the Machine Learning. Morgan Kaufmann Publishers Inc., Bari, Italy.
Fu, B., Hower, J.C., Zhang, W., Luo, G., Hu, H., Yao, H., 2022. A review of rare earth elements and yttrium in coal ash: Content, modes of occurrences, combustion behavior, and extraction methods. Prog. Energy Combust. Sci. 88, 100954 https://doi.org/10.1016/j.pecs.2021.100954.
Gaustad, G., Williams, E., Leader, A., 2021. Rare earth metals from secondary sources: Review of potential supply from waste and byproducts. Resour. Conserv. Recycl. 167, 105213 https://doi.org/10.1016/j.resconrec.2020.105213.
Geological Survey, U.S., 2022. Mineral Commodity Summaries 2022. U.S. Geological Survey, p. 202. https://doi.org/10.3133/mcs2022.
Goode, J.R., 2023. Magnet rare earth production must double in ten years. Where will production come from?. In: Proceedings of the 61st Conference of Metallurgists, COM 2022. Springer, Cham. https://doi.org/10.1007/978-3-031-17425-4_87.
Gregory, D.D., Cracknell, M.J., Large, R.R., McGoldrick, P., Kuhn, S., Maslennikov, V.V., Baker, M.J., Fox, N., Belousov, I., Figueroa, M.C., Steadman, J.A., Fabris, A.J., Lyons, T.W., 2019. Distinguishing ore deposit type and barren sedimentary pyrite using laser ablation-inductively coupled plasma-mass spectrometry trace element data and statistical analysis of large data sets. Econ. Geol. 114, 771–786. https://doi.org/10.5382/econgeo.4654.
Gromet, L.P., Haskin, L.A., Korotev, R.L., Dymek, R.F., 1984. The "North American shale composite": its compilation, major and trace element characteristics. Geochim. Cosmochim. Acta 48, 2469–2482. https://doi.org/10.1016/0016-7037(84)90298-9.
Grunsky, E.C., de Caritat, P., 2020. State-of-the-art analysis of geochemical data for mineral exploration. In: Geochemistry: Exploration, Environment, Analysis, 20, pp. 217–232. https://doi.org/10.1144/geochem2019-031.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer Series in Statistics. Springer, New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7.
Hein, J., Koschinsky, A., Mikesell, M., Mizell, K., Glenn, C., Wood, R., 2016. Marine phosphorites as potential resources for heavy rare earth elements and yttrium. Minerals 6, 88. https://doi.org/10.3390/min6030088.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67. https://doi.org/10.1080/00401706.1970.10488634.
Hu, B., Zeng, L.-P., Liao, W., Wen, G., Hu, H., Li, M.Y.H., Zhao, X.-F., 2022. The origin and discrimination of high-ti magnetite in magmatic-hydrothermal systems: insight from machine learning analysis. Econ. Geol. 117, 1613–1627. https://doi.org/10.5382/econgeo.4946.
Hu, Y., Ma, J., Wang, J., Niu, H., Yang, Z., Hao, H., Panchal, B., 2023. Differentiation of rare earth elements in coal combustion products from the Handan Power Plant, Hebei Province. China. Sustainability 15, 3420. https://doi.org/10.3390/su15043420.
Huang, Z., Fan, M., Tian, H., 2020. Rare earth elements of fly ash from Wyoming's Powder River Basin coal. J. Rare Earths 38, 219–226. https://doi.org/10.1016/j.jre.2019.05.004.
Jensen, G.K.S., Pollard, A., Rostron, B.J., 2020. Lithium Concentration in the Duperow Formation: Preliminary Results of Geochemical Analysis of Core Samples from Two Wells in Southeastern Saskatchewan. in Summary of Investigations 2020, Volume 1, Saskatchewan Geological Survey, Saskatchewan Ministry of Energy and Resources, Miscellaneous Report 2020-4.1, Paper A-2. Regina, SK.
Johannesson, K.H., Hawkins, D.L., Cortés, A., 2006. Do Archean chemical sediments record ancient seawater rare earth element patterns? Geochim. Cosmochim. Acta 70, 871–890. https://doi.org/10.1016/j.gca.2005.10.013.
Jolliffe, I., 2002. Principal Component Analysis, 2nd edition. Springer, New York.
Jowitt, S.M., 2022. Mineral economics of the rare-earth elements. Mater. Res. Soc. Bull. 47, 1–7. https://doi.org/10.1557/s43577-022-00289-3.
Kamber, B.S., Webb, G.E., Gallagher, M., 2014. The rare earth element signal in Archaean microbial carbonate: information on ocean redox and biogenicity. J. Geol. Soc. Lond. 171, 745–763. https://doi.org/10.1144/jgs2013-110.
Kaplan, U.E., Topal, E., 2020. A new ore grade estimation using combine machine learning algorithms. Minerals 10, 847. https://doi.org/10.3390/min10100847.
Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2019. Machine learning for the geosciences: challenges and opportunities. IEEE Trans. Knowl. Data Eng. 31, 1544–1554. https://doi.org/10.1109/TKDE.2018.2861006.
Kassambara, A., Mundt, F., 2020. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. https://CRAN.R-project.org/package=factoextra.
Kato, Y., Fujinaga, K., Nakamura, K., Takaya, Y., Kitamura, K., Ohta, J., Toda, R., Nakashima, T., Iwamori, H., 2011. Deep-sea mud in the Pacific Ocean as a potential resource for rare-earth elements. Nat. Geosci. 4, 535–539. https://doi.org/10.1038/ngeo1185.
Kokh, M.A., Luais, B., Truche, L., Boiron, M.C., Peiffert, C., Schumacher, A., 2021. Quantitative measurement of rare earth elements in brines: isolation from the charged matrix versus direct LA-ICP-MS measurements – a comparative study. Geostand. Geoanal. Res. 45, 341–358. https://doi.org/10.1111/ggr.12376.
Kolker, A., Scott, C., Hower, J.C., Vazquez, J.A., Lopano, C.L., Dai, S., 2017. Distribution of rare earth elements in coal combustion fly ash, determined by SHRIMP-RG ion microprobe. Int. J. Coal Geol. 184, 1–10. https://doi.org/10.1016/j.coal.2017.10.002.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3.
Lange, C.N., Camargo, I.M.C., Figueiredo, A.M.G.M., Castro, L., Vasconcellos, M.B.A., Ticianelli, R.B., 2017. A Brazilian coal fly ash as a potential source of rare earth elements. J. Radioanal. Nucl. Chem. 311, 1235–1241. https://doi.org/10.1007/s10967-016-5026-8.
Lawley, C.J.M., McCafferty, A.E., Graham, G.E., Huston, D.L., Kelley, K.D., Czarnota, K., Paradis, S., Peter, J.M., Hayward, N., Barlow, M., Emsbo, P., Coyan, J., San Juan, C.A., Gadd, M.G., 2022. Data–driven prospectivity modelling of sediment–hosted

Zn–Pb mineral systems and their critical raw materials. Ore Geol. Rev. 141, 104635 https://doi.org/10.1016/j.oregeorev.2021.104635.
Lee, J., Bazilian, M., Sovacool, B., Hund, K., Jowitt, S.M., Nguyen, T.P., Månberger, A., Kah, M., Greene, S., Galeazzi, C., Awuah-Offei, K., Moats, M., Tilton, J., Kukoda, S., 2020. Reviewing the material and metal security of low-carbon energy transitions. Renew. Sust. Energ. Rev. 124, 109789 https://doi.org/10.1016/j.rser.2020.109789.
Levin, L.A., Amon, D.J., Lily, H., 2020. Challenges to the sustainability of deep-seabed mining. Nat. Sustain. 3, 784–794. https://doi.org/10.1038/s41893-020-0558-x.
Lewis, A.J., Komninou, A., Yardley, B.W.D., Palmer, M.R., 1998. Rare earth element speciation in geothermal fluids from Yellowstone National Park, Wyoming, USA. Geochim. Cosmochim. Acta 62, 657–663. https://doi.org/10.1016/S0016-7037(97)00367-0.
Lindsay, J.J., Hughes, H.S.R., Yeomans, C.M., Andersen, J.C.Ø., McDonald, I., 2021. A machine learning approach for regional geochemical data: Platinum-group element geochemistry vs geodynamic settings of the North Atlantic Igneous Province. Geosci. Front. 12, 101098 https://doi.org/10.1016/j.gsf.2020.10.005.
Linnen, R.L., Samson, I.M., Williams-Jones, A.E., Chakhmouradian, A.R., 2014. Geochemistry of the rare-earth element, Nb, Ta, Hf, and Zr deposits. In: Treatise on Geochemistry. Elsevier, pp. 543–568. https://doi.org/10.1016/B978-0-08-095975-7.01124-4.
Liu, H., Guo, H., Xing, L., Zhan, Y., Li, F., Shao, J., Niu, H., Liang, X., Li, C., 2016. Geochemical behaviors of rare earth elements in groundwater along a flow path in the North China Plain. J. Asian Earth Sci. 117, 33–51. https://doi.org/10.1016/j.jseaes.2015.11.021.
Lopez, G.P., Rokosh, C.D., Weiss, J.A., Pawlowicz, J.G., 2020. Inorganic geochemistry of bulk samples from selected Alberta geological units (tabular data, tab-delimited format). In: Alberta Energy Regulator/Alberta Geological Survey, AER/AGS Digital Data 2019-0021.
Mänd, K., Lalonde, S.V., Paiste, K., Thoby, M., Lumiste, K., Robbins, L.J., Kreitsmann, T., Romashkin, A.E., Kirsimäe, K., Lepland, A., Konhauser, K.O., 2021. Iron isotopes reveal a benthic iron shuttle in the palaeoproterozoic zaonega formation: basinal restriction, euxinia, and the effect on global palaeoredox proxies. Minerals 11, 368. https://doi.org/10.3390/min11040368.
Mastalerz, M., Drobniak, A., Branam, T., 2022. Coal and coal byproducts as potential sources of rare earth elements (REE) in Indiana. Indiana J. Earth Sci. 4.
McArthur, J.M., Walsh, J.N., 1984. Rare-earth geochemistry of phosphorites. Chem. Geol. 47, 191–220. https://doi.org/10.1016/0009-2541(84)90126-8.
Miranda, M.A., Ghosh, A., Mahmodi, G., Xie, S., Shaw, M., Kim, S., Krzmarzick, M.J., Lampert, D.J., Aichele, C.P., 2022. Treatment and recovery of high-value elements from produced water. Water 14, 880. https://doi.org/10.3390/w14060880.
Mirza, T.A., Karim, K.H., Ridha, S.M., Fatah, C.M., 2021. Major, trace, rare earth element, and stable isotope analyses of the Triassic carbonates along the northeastern Arabian Plate margin: a key to understanding paleotectonics and paleoenvironment of the Avroman (Biston) limestone formation from Kurdistan region, northeastern Iraq. Carbonates Evaporites 36, 66. https://doi.org/10.1007/s13146-021-00733-6.
Modi, P., Jamal, A., Singh, N., 2021. Coal characterization and occurrence of rare earth elements in coal and coal-ash of Sohagpur Coalfield, Madhya Pradesh, India. Int. J. Coal Prep. Util. 1–14 https://doi.org/10.1080/19392699.2021.1923489.
Möller, P., Dulski, P., De Lucia, M., 2021. Rey patterns and their natural anomalies in waters and brines: the correlation of Gd and Y anomalies. Hydrology 8. https://doi.org/10.3390/hydrology8030116.
Montross, S.N., Bagdonas, D., Paronish, T., Bean, A., Gordon, A., Creason, C.G., Thomas, B., Phillips, E., Britton, J., Quillian, S., Rose, K., 2022. On a unified core characterization methodology to support the systematic assessment of rare earth elements and critical minerals bearing unconventional carbon ores and sedimentary strata. Minerals 12, 1159. https://doi.org/10.3390/min12091159.
Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorobot. 7 https://doi.org/10.3389/fnbot.2013.00021.
Natural Resources Canada, 2022. The Canadian Critical Minerals Strategy. https://www.canada.ca/content/dam/nrcan-rncan/site/critical-minerals/Critical-minerals-strategyDec09.pdf. Retrieved August 29, 2023.
Nothdurft, L.D., Webb, G.E., Kamber, B.S., 2004. Rare earth element geochemistry of late Devonian reefal carbonates, Canning Basin, Western Australia: confirmation of a seawater REE proxy in ancient limestones. Geochim. Cosmochim. Acta 68, 263–283. https://doi.org/10.1016/S0016-7037(03)00422-8.
Ogen, Y., Denk, M., Glaesser, C., Eichstaedt, H., 2022. A novel method for predicting the geochemical composition of tailings with laboratory field and hyperspectral airborne data using a regression and classification-based approach. Eur. J. Remote Sens. 55, 453–470. https://doi.org/10.1080/22797254.2022.2104173.
Otero, N., Tolosana-Delgado, R., Soler, A., Pawlowsky-Glahn, V., Canals, A., 2005. Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a Mediterranean river. Water Res. 39, 1404–1414. https://doi.org/10.1016/j.watres.2005.01.012.
Parsa, M., Lentz, D.R., Walker, J.A., 2022. Predictive Modeling of Prospectivity for VHMS Mineral Deposits, Northeastern Bathurst Mining Camp, NB, Canada, Using an Ensemble Regularization Technique. Natural Resources Research. https://doi.org/10.1007/s11053-022-10133-9.
Pawlowsky-Glahn, V., Egozcue, J.J., 2006. Compositional data and their analysis: an introduction. Geol. Soc. Spec. Publ. 264, 1–10. https://doi.org/10.1144/GSL.SP.2006.264.01.01.
Quillinan, S., Nye, C., Engle, M., Bartos, T., Brant, J., Bagdonas, D., McLing, T., McLaughlin, J.F., 2018. Assessing Rare Earth Element Concentrations in Geothermal and Oil and Gas Produced Waters: A Potential Domestic Source of Strategic Mineral Commodities (No. DE-EE0007603). US Department of Energy.
R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Reimann, C., Filzmoser, P., Garrett, R.G., Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R, 1st ed. Wiley. https://doi.org/10.1002/9780470987605.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001.
Rodriguez-Galiano, V.F., Chica-Rivas, M., 2014. Evaluation of different machine learning methods for land cover mapping of a Mediterranean area using multi-seasonal Landsat images and Digital Terrain Models. Int. J. Digit. Earth 7, 492–509. https://doi.org/10.1080/17538947.2012.748848.
Rokosh, C.D., Crocq, C.S., Pawlowicz, J.G., Brazzoni, T., 2016. Inorganic geochemistry of Alberta geological units for shale- and siltstone-hosted hydrocarbon evaluation (tabular data, tab-delimited format). In: Alberta Energy Regulator, AER/AGS Digital Data 2016-0001.
Roth, E., Bank, T, Howard, B., Granite, E., 2017. Rare earth elements in Alberta oil sand process streams. Energy Fuel 31, 4714–4720. https://doi.org/10.1021/acs.energyfuels.6b03184.
Sandeep, P., Maity, S., Mishra, S., Chaudhary, D.K., Dusane, C.B., Pillai, A.S., Kumar, A. V., 2023. Estimation of rare earth elements in Indian coal fly ashes for recovery feasibility as a secondary source. J. Hazard. Mater. Adv. 10, 100257 https://doi.org/10.1016/j.hazadv.2023.100257.
Schnitzler, N., Ross, P.S., Gloaguen, E., 2019. Using machine learning to estimate a key missing geochemical variable in mining exploration: application of the random forest algorithm to multi-sensor core logging data. J. Geochem. Explor. 205, 106344 https://doi.org/10.1016/j.gexplo.2019.106344.
Seredin, V.V., Dai, S., 2012. Coal deposits as potential alternative sources for lanthanides and yttrium. Int. J. Coal Geol. 94, 67–93. https://doi.org/10.1016/j.coal.2011.11.001.
Smith, Y., Kumar, P., McLennan, J., 2017. On the extraction of rare earth elements from geothermal brines. Resources 6, 39. https://doi.org/10.3390/resources6030039.
Song, Y., Liang, J., Lu, J., Zhao, X., 2017. An efficient instance selection algorithm for k nearest neighbor regression. Neurocomputing 251, 26–34. https://doi.org/10.1016/j.neucom.2017.04.018.
Taggart, R.K., Hower, J.C., Dwyer, G.S., Hsu-Kim, H., 2016. Trends in the rare earth element content of U.S.-based coal combustion fly ashes. Environ. Sci. Technol. 50, 5919–5926. https://doi.org/10.1021/acs.est.6b00085.
Takaya, Y., Yasukawa, K., Kawasaki, T., Fujinaga, K., Ohta, J., Usui, Y., Nakamura, K., Kimura, J.-I., Chang, Q., Hamada, M., Dodbiba, G., Nozaki, T., Iijima, K., Morisawa, T., Kuwahara, T., Ishida, Y., Ichimura, T., Kitazume, M., Fujita, T., Kato, Y., 2018. The tremendous potential of deep-sea mud as a source of rare-earth elements. Sci. Rep. 8, 5763. https://doi.org/10.1038/s41598-018-23948-5.
Taylor, S.R., McLennan, S.M., 1985. The Continental Crust: Its Composition and Evolution. Blackwell, Oxford, pp. 1–312.
Templ, M., Filzmoser, P., Reimann, C., 2008. Cluster analysis applied to regional geochemical data: problems and possibilities. Appl. Geochem. 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. John Wiley and Sons, ISBN 978-0-470-71135-4. https://doi.org/10.1002/9781119976462.ch25.
Van Den Boogaart, K.G., Tolosana-Delgado, R., 2008. compositions": A unified R package to analyze compositional data. Comput. Geosci. 34, 320–338. https://doi.org/10.1016/j.cageo.2006.11.017.
Vural, A., 2020. Investigation of the relationship between rare earth elements, trace elements, and major oxides in soil geochemistry. Environ. Monit. Assess. 192, 124. https://doi.org/10.1007/s10661-020-8069-9.
Wagner, N.J., Matiane, A., 2018. Rare earth elements in select Main Karoo Basin (South Africa) coal and coal ash samples. Int. J. Coal Geol. 196, 82–92. https://doi.org/10.1016/j.coal.2018.06.020.
Wang, Xiaomei, Wang, Xiaoming, Pan, Z., Yin, X., Chai, P., Pan, S., Yang, Q., 2019. Abundance and distribution pattern of rare earth elements and yttrium in vitrain band of high-rank coal from the Qinshui basin, northern China. Fuel 248, 93–103. https://doi.org/10.1016/j.fuel.2019.03.054.
Wei, T., Simko, V., 2021. R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92). https://github.com/taiyun/corrplot.
Wu, L., Ma, L., Huang, G., Li, J., Xu, H., 2022. Distribution and speciation of rare earth elements in coal fly ash from the Qianxi Power Plant, Guizhou province, southwest China. Minerals 12, 1089. https://doi.org/10.3390/min12091089.
Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., Tu, M., 2018. Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances. J. Pet. Sci. Eng. 160, 182–193. https://doi.org/10.1016/j.petrol.2017.10.028.
Yin, X., Martineau, C., Demers, I., Basiliko, N., Fenton, N.J., 2021. The potential environmental risks associated with the development of rare earth element production in Canada. Environ. Rev. 29, 354–377. https://doi.org/10.1139/er-2020-0115.
Zhang, S., Dai, S., Finkelman, R.B., Graham, I.T., French, D., Hower, J.C., Li, X., 2019. Leaching characteristics of alkaline coal combustion by-products: a case study from a coal-fired power plant, Hebei Province, China. Fuel 255, 115710. https://doi.org/10.1016/j.fuel.2019.115710.

Zhou, S., Zhou, K., Wang, J., Yang, G., Wang, S., 2018. Application of cluster analysis to geochemical compositional data for identifying ore-related geochemical anomalies. Front. Earth Sci. 12, 491–505. https://doi.org/10.1007/s11707-017-0682-8.

Zhu, J.-J., Yang, M., Ren, Z.J., 2023. Machine learning in environmental research: common pitfalls and best practices. Environ. Sci. Technol. 57, 17671–17689. https://doi.org/10.1021/acs.est.3c00026.

Zuo, R., Carranza, E.J.M., 2011. Support vector machine: a tool for mapping mineral prospectivity. Comput. Geosci. 37, 1967–1975. https://doi.org/10.1016/j.cageo.2010.09.014.