Artificial Intelligence in Geosciences 2 (2021) 128–147

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences

Towards a fully data-driven prospectivity mapping methodology: A case study of the Southeastern Churchill Province, Québec and Labrador

Steven E. Zhang [a], Julie E. Bourdeau [b], Glen T. Nwaila [c,*], David Corrigan [d]

[a] SmartMin Limited, 39 Kiewiet Street, Helikon Park, 1759, South Africa
[b] Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, K1A 0E8, Canada
[c] School of Geosciences, University of the Witwatersrand, 1 Jan Smuts Ave, Johannesburg, 2000, South Africa
[d] 79 Lynn Street, Gatineau, Québec, J9H 1B5, Canada

ARTICLE INFO

Keywords:
ML
Mineral prospectivity mapping
Principal component analysis
Geochemical anomaly
REEs

ABSTRACT

Mineral exploration campaigns are financially risky. Several state-of-the-art methods have been developed to mitigate the risk, including predictive modelling of mineral prospectivity using principal component analysis (PCA) and geographic information systems (GIS). The PCA and GIS approach is currently considered acceptable for generating mineral exploration targets. However, some of its limitations are the dependence on sample stoichiometry (e.g., the existence of minerals), the necessity of log-ratio transformations when dealing with compositional data, and manual interpretation and use of principal components to enhance potential geochemical anomalies for prospectivity mapping. In this study, we generalize the fundamental ideas behind the PCA and GIS approach by developing a new data-driven approach using ML. We showcase a new workflow capable of generating either intermediate evidence layers or final prospectivity maps that depict major regional geochemical anomalies using multi-element geochemical data from Southeastern Churchill Province (Québec and Labrador), Canada. The region is known for its REEs endowment and the data were gathered for prospectivity mapping. A comparison with the established multivariate hybrid data- and knowledge-based approach revealed that on a roughly comparable basis of the amount of manual effort, our new data-driven procedure can much more accurately identify geochemical anomalies in both univariate and multivariate applications. The results of our prospectivity mapping corroborate with the ground truth or known geological anomalies in the studied region. These findings have potentially wider implications on exploration target generation, where project risks (financial, environmental, political, etc.) and geochemical anomalies must be quantified using robust and effective data-driven approaches. In addition, our methodology is more replicable and objective, as manual geoscientific interpretation is not required during the detection of geochemical anomalies.

1. Introduction

Prospectivity mapping transforms geoscientific data into maps that depict regional potential or favourability of deposits or their proxies (e.g., indicator elements) using some combination of knowledge- and data-driven methods (e.g., Chung and Agterberg, 1980; Bonham-Carter, 1994; Harris and Pan, 1999; Wright and Bonham-Carter, 1996; Brown et al., 2000; Carranza et al., 2008; Harris et al., 2015; Grunsky and de Caritat, 2019). A key outcome is to delineate interesting regions (e.g., anomalies) that can be preferentially targeted for further investigation and interpretation. Knowledge-driven approaches use heuristic models of discipline-specific knowledge, such as characteristics of a deposit being sought to guide the use of data (e.g., geochemical and geophysical maps). Data-driven approaches often makes use of existing ground truth to train exploration models (Bonham-Carter, 1994; Wright and Bonham-Carter, 1996; Carranza, 2008, 2009a, b), which renders such methods more suitable for brownfield than greenfield exploration. However, the discovery of spatially-coherent geochemical anomalies is a type of anomaly detection task that can occur in some combination of the variable and the spatial space using unsupervised machine learning, which does not require ground truth. A data-driven technique based on multivariate statistical methods to discover anomalous processes exists (e.g., Carranza, 2008; Grunsky and de Caritat, 2019). However, it requires manual interpretation and therefore relies on some discipline knowledge. This

Abbreviations: ML, Machine Learning; REEs, Rare Earth Elements.

* Corresponding author.
E-mail address: glen.nwaila@wits.ac.za (G.T. Nwaila).

https://doi.org/10.1016/j.aiig.2022.02.002
Received 8 December 2021; Received in revised form 16 February 2022; Accepted 17 February 2022
Available online 1 March 2022

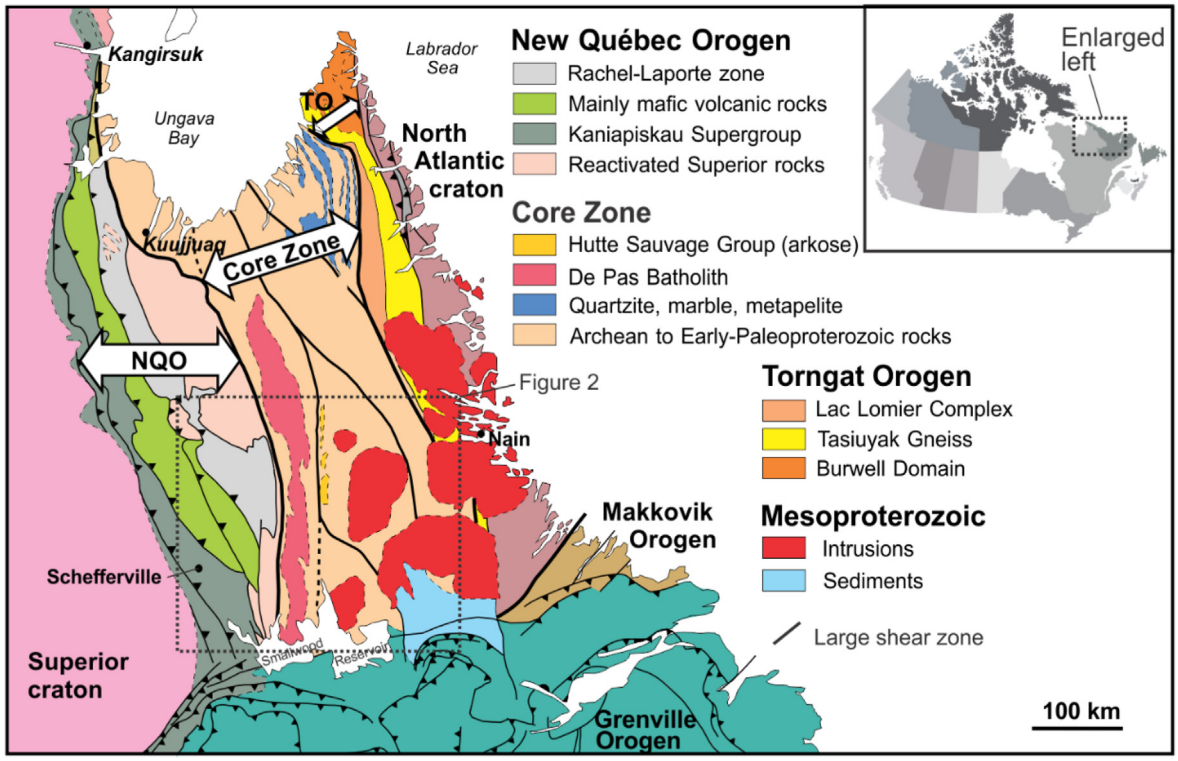

Fig. 1. Location of and simplified geological map of the Southeastern Churchill Province. Abbreviations: New Québec Orogen (NQO); Torngat Orogen (TO). Location and extent of the study area (Fig. 2) is shown by the box. Image modified after James and Dunning (2000), and Corrigan et al. (2018).

method ("the multivariate method", e.g., Carranza, 2008; Grunsky and de Caritat, 2019) may be generalized using machine learning methods that weakens the technique's reliance on geoscientific knowledge and manual interpretation. In this paper, we demonstrate a generalization of the variable- and geospatial-space based anomaly detection using compositional geochemical data that removes some key weaknesses of the multivariate method and does not require ground truth (of anomalies). Therefore, our technique could be used for either greenfield or brownfield exploration.

Exploration geochemical data contains spatial information and sample chemistry (e.g., of bedrock, glacial till, water, soil, stream and lake sediment) and is intended to capture the effects of geological processes that may control elemental distribution, such as lithological variability, metamorphic or hydrothermal overprinting and other secondary processes. Some processes may be responsible for chemical anomalies that are associated with the presence of mineral deposits (Grunsky and de Caritat, 2019). Where sample chemistry is multivariate in the form of compositional data, concentrations are reported as relative proportions. A data-driven technique can identify geochemical anomalies using compositional data (Harris et al., 2015; Chen et al., 2018; Grunsky and de Caritat, 2019; Grunsky and Arne, 2020). However, the method as summarized by Grunsky and de Caritat (2019) exhibits some key weaknesses. The first is the assumption of geochemistry as a proxy for mineralogy through elemental stoichiometry, which may be unmet in some cases (e.g., volcanic or clay-rich rocks and/or in highly altered areas, see Zhang et al., 2021). As the strength of mineralogy-controlled variability weakens, the ability to produce useful and interpretable multivariate elemental associations is gradually lost. Although this does not nullify the technique in such situations, it does increase the burden of manual interpretation and validation. The second is the use of manual geological and geochemical interpretation of key stoichiometries, which increases mapping subjectivity and renders this method more knowledge-driven (Carranza, 2008; Harris et al., 2015, Chen et al., 2018; Grunsky and de Caritat, 2019). This results from a reliance on principal component analysis (PCA), in a manner that is more resemblant of multivariate statistical analysis than machine learning (ML; e.g., Carranza, 2008; Zuo, 2011; Harris et al., 2015; Grunsky and de Caritat, 2019). For data-driven approaches, PCA serves two main purposes: (1) a reduction of feature space dimensionality, and (2) an automation of the extraction of multivariate relationships. Where PCA is applied to geochemical data to extract multivariate relationships, the interpretation of principal components is unavoidable, in order to discriminate between background processes and anomalous (e.g., deposit-forming) processes (Carranza, 2008; Zuo, 2011; Grunsky et al., 2014; Harris et al., 2015; Grunsky and de Caritat, 2019). In contrast to this method (e.g., Grunsky and de Caritat., 2019), there exists a collection of deep learning-based approaches to the generation of prospectivity maps (Zuo et al., 2019 and references therein). Such methods use highly abstract and powerful neural networks to extract patterns that may be indicative of relevant geological processes. These approaches are perhaps the highest in performance (exceeding human performance, Zuo et al., 2019) and as such, are at the leading-edge of data-driven applications in many fields (e.g., He et al., 2015 and Lundervold and Lundervold, 2019). However, they produce models that are inherently complex and difficult to interpret (Zuo et al., 2019) and in response to this, an entire field of explainable artificial intelligence exists and is driven by the need to be able to understand decisions (e.g., Linardatos et al., 2021). For exploration, the uncertainty of prospectivity maps is generally an unsolved problem, which can be compounded by the lack of explain-ability of complex ML models (e.g., neural networks and ensembles). However, it is possible to simultaneously leverage desirable aspects of algorithmically simpler approaches that invoke some in-discipline knowledge (Grunsky and de Caritat, 2019), while accommodating complex and powerful algorithms (e.g., Luo et al., 2020 and Zuo et al., 2019 and references therein). To this effect, we develop and generalize the Grunsky and de Caritat (2019) technique using ML to automatically delineate geochemical anomalies in a manner that both overcomes the method's major weaknesses and explicitly make provisions for the use of many ML algorithms. The latter

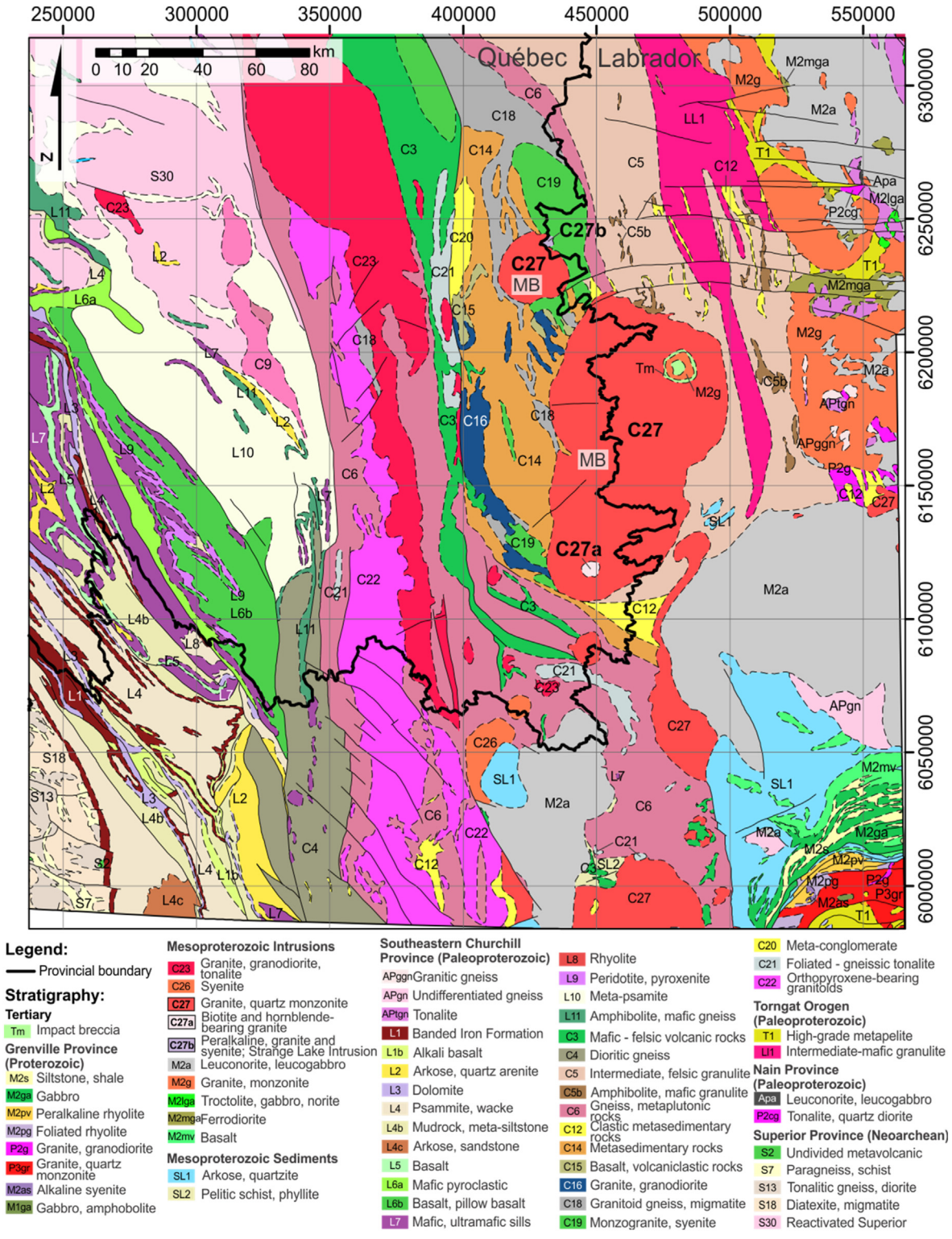

Fig. 2. Simplified bedrock geology map of the study area (UTM Zone 20, NAD 83). Lake sediment samples used in this study are all located in Labrador (NTS 13-L, 13-M, 14-D, 23-I and 23-J). The Mistastin Batholith (labelled MB) is located at the provincial boundary between Québec and Labrador (near center of the map).

improvement allows our technique to offer a choice of explain-ability and performance. To demonstrate the benefits of our technique relative to a control method, we provide a qualitative comparison between geochemical maps that are produced using our method and the Grunsky and de Caritat (2019) multivariate approach. In addition to our technique generating more selective and less noisy prospectivity maps, we also demonstrate that our approach is more data-driven, more automatable, is extensible to complex geological environments, does not require data transformations and imputation and can be implemented for non-stoichiometric compositions.

2. Background and regional geology

Our data resulted from lake sediment sampling of the Southeastern

Table 1
Model parameters used in the grid search.

Algorithm	Parameter Grid
kNN	k = {1, 3, 5, 7, 9, 11}
SVM	C = {10, 100, 250, 500, 750, 1000}, ε = {0.01, 0.1, 0.5, 1.0}, kernel = {linear, RBF}
Elastic net	ρ = {0.1, 0.25, 0.5, 0.75, 1.0}
Random forest	Ensemble size = 500; maximum depth = {5, 4, 3, 2, 1, unlimited}, maximum number of features = {1, 2, 3, 4, 5}, minimum number of samples for a split = {2, 3, 4}, minimum number of samples for a leaf = {1, 2, 3}
AdaBoost	Number of classifiers = {50, 250, 500}, base algorithm = decision tree with the same parameter grid as the random forest algorithm
ANN	α = {0.001, 0.01, 0.1, 1.0}, activation = {identity, logistic, tanh, relu}, learning rate = {constant, inverse scaling, adaptive}

Churchill Province, which is at the border between northern Québec and Labrador (Fig. 1). It is a glaciated landscape that is under exploration for rare earth elements (REEs), which centers around the Strange Lake peralkaline complex (Fig. 2). Anomalies in lake sediments and lake waters as well as gamma-ray spectrometry led to the discovery of the Strange Lake deposit in 1980 (Friske et al., 1996a, b; Zajac, 2015), which is one of the world's largest deposits of Zr, Y and heavy REEs (Zajac, 2015). REEs are critical to modern societies due to their role in the high-tech, transportation and energy sectors (Balaram, 2019), and the need to diversify supply chains to mitigate geopolitical instability (Balaram, 2019; Overland, 2019). For these reasons, the European Union, United States of America, Australia, People's Republic of China and Canada have enacted legislation and policies to secure their supply chains of REEs (European Commission, 2011, 2014, 2017a, 2017b, 2020; Ghorbani, 2018; U.S. Geological Survey, 2018; Rizzo et al., 2020; Natural Resources Canada, 2021). Prior knowledge of the Strange Lake deposit provides the necessary, albeit qualitative ground truth for the verification of geochemical anomaly maps.

The Southeastern Churchill Province is an elongated sliver of Archean to early-Paleoproterozoic continental crust that was accreted during the Paleoproterozoic, through collision of the Superior (to the west) and North Atlantic (to the east) cratons (Fig. 1; James et al., 1996; James and Dunning, 2000; Wardle et al., 2002; Hammouche et al., 2012; Corrigan et al., 2018). To the west, the southeastern Churchill Province is flanked by the New Québec Orogen (1.83–1.80 Ga; Hoffman, 1988; Perreault and Hynes, 1990; Moorhead and Hynes, 1990) and to the east by the Torngat Orogen (1.87–1.86 Ga; Wardle et al., 2002). The extent to which these orogens affected the Core Zone (Fig. 1) is relatively unknown, although the New Québec Orogen overprint appears to be more widespread (Corrigan et al., 2018). As a result of the accretion, numerous folds and large shear zones transect the region (Fig. 1). Rocks in the area consist mostly of paragneiss, orthogneiss, migmatite and low-to medium--metamorphic grade supracrustal and plutonic rocks that were intruded by post-orogen, Mesoproterozoic-age intrusions (Hammouche et al., 2012; Corrigan et al., 2018).

The Mistastin Batholith (C27 in Fig. 2; 1400 Ga), which outcrops as two semi-circular intrusions covering an area of 5600 km^2, is of particular importance in the study area (Hammouche et al., 2012). The largest of the intrusions by planar dimensions (to the south) is ellipsoid-shaped (long axis striking at N20E) and measures approximately 144 by 65 km. To the north is a smaller, circularly-shaped intrusion, measuring approximately 26 by 27 km. The Mistastin Batholith is chiefly composed of intermediate potassic intrusive rocks with lesser amounts of monzonite, quartz syenite, granite, anorthosite, leucogabbronorite and leucogabbro (Van der Leeden, 1995). Notably, this batholith is locally highly enriched in REEs, Zr, Y, ± Be, Nb, U and Th (Hammouche et al., 2012). One of these zones of enrichment is hosted by the Misery Lake

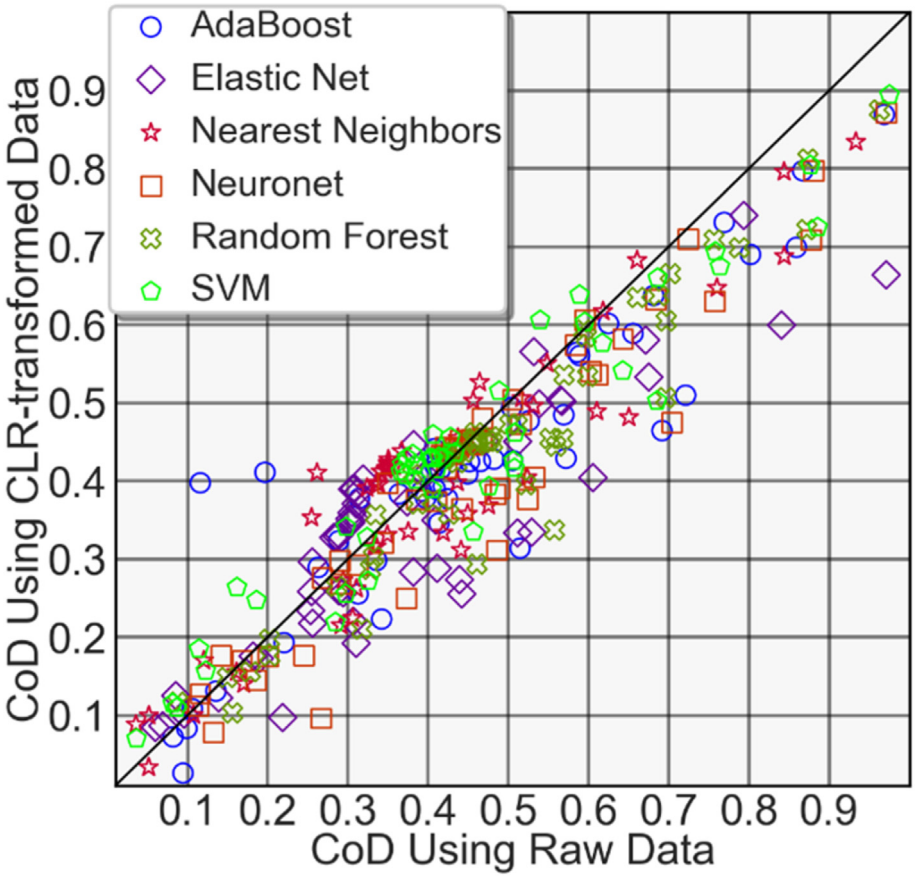

Fig. 3. Comparison of prediction performance using the CoD metric for all elements.

Table 2
Selected best hyperparameters for the elements Y and Rb for all algorithms. Raw = non-transformed data; CLR = CLR-transformed data.

Element	Data type	Method	CoD	Parameters
Y	Raw	kNN	0.358	$k = 11$
Y	Raw	Elastic net	0.303	$\rho = 1.0$
Y	Raw	SVM	0.373	$C = 1000$, ε: 1.0, kernel = RBF
Y	Raw	Random forest	0.455	Maximum depth = 0, maximum number of features = 4, minimum number of samples for a leaf = 2, minimum number of samples for a split = 3
Y	Raw	AdaBoost	0.419	Maximum depth = unlimited, minimum number of samples per split = 2, minimum number of samples for a leaf = 1, number of classifiers = 250
Y	Raw	ANN	0.433	Activation = tanh, $\alpha = 0.01$, learning rate: constant
Y	CLR	kNN	0.419	$k = 11$
Y	CLR	Elastic net	0.346	$\rho = 0.75$
Y	CLR	SVM	0.429	$C = 500$, ε: 1.0, kernel: RBF
Y	CLR	Random forest	0.453	Maximum depth = 0, maximum number of features = 3, minimum number of samples for a leaf = 3, minimum number of samples for a split = 2
Y	CLR	AdaBoost	0.393	Maximum depth = unlimited, maximum number of features per split = 3, minimum number of samples for a leaf = 3, minimum number of classifiers = 250
Y	CLR	ANN	0.441	Activation = tanh, $\alpha = 0.01$, learning rate: constant
Rb	Raw	kNN	0.933	$k = 5$
Rb	Raw	Elastic net	0.971	$\rho = 1.0$
Rb	Raw	SVM	0.975	$C = 1000$, ε: 0.1, kernel: RBF
Rb	Raw	Random forest	0.962	Maximum depth = 0, maximum number of features = 4, minimum number of samples for a leaf = 2, minimum number of samples for a split = 4
Rb	Raw	AdaBoost	0.969	Maximum depth = unlimited, maximum number of features per split = 4, minimum number of samples for a leaf = 2, minimum number of classifiers = 250
Rb	Raw	ANN	0.971	Activation = tanh, $\alpha = 0.001$, learning rate: constant
Rb	CLR	kNN	0.834	$k = 5$
Rb	CLR	Elastic net	0.664	$\rho = 0.75$
Rb	CLR	SVM	0.894	$C = 1000$, ε: 0.5, kernel: RBF
Rb	CLR	Random forest	0.875	Maximum depth = 0, maximum number of features = 4, minimum number of samples for a leaf = 2, minimum number of samples for a split = 2
Rb	CLR	AdaBoost	0.869	Maximum depth = unlimited, maximum number of features per split = 4, minimum number of samples for a leaf = 2, minimum number of classifiers = 250
Rb	CLR	ANN	0.871	Activation = tanh, $\alpha = 0.001$, learning rate: constant

peralkaline syenite (C27a in Fig. 2; 1409.7 ± 1.2 Ma), located in the southern portion of the largest intrusion (David et al., 2012). The most notable deposit known is hosted by the Strange Lake peralkaline complex (C27b in Fig. 2; 1240 ± 2Ma) with indicated mineral resources of 278 Mt at 0.93% total REE oxide (Miller, 1990; Gowans et al., 2014; Zajac, 2015).

The batholith is highly weathered, likely due to intensive hydrothermal alteration in zones associated with REE-mineralization prior to Quaternary glaciation events. Glacial erosion by the Laurentide Ice Sheet throughout the Wisconsin glaciation remobilized the hydrothermally altered bedrocks and along with it, large volumes of REE-rich debris towards the northeast. Lake sediment and lake water geochemistry, boulder mapping, matrix till geochemistry, indicator minerals and airborne gamma-ray spectroscopy data reveal that the dispersal trend can be detected in excess of 50 km down ice (Geological Survey of Canada, 1980; Batterson, 1989; Zajac, 2015; Paulen et al., 2017; McClenaghan et al., 2017, 2019).

3. Methods

3.1. Source data

The lake sediment geochemical data used for this study, the sampling protocol, analytical methodology, quality control and quality assurance procedures are documented in McCurdy et al. (2016). The survey consists of 3441 samples, including field duplicates, and was sampled at a resolution of about one sample per 13 km^2. The data used in this study is from a campaign to re-analyze lake sediment samples using modern analytical methods as part of the Geological Survey of Canada's GEM program (GEM, 2019). Geochemical analyses were conducted at Bureau Veritas in Vancouver, Canada. A total of 65 elements were analyzed: Ag, Al, As, Au, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Gd, Ge, Hf, Hg, Ho, In, K, La, Li, Lu, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Pd, Pr, Pt, Rb, Re, S, Sb, Sc, Se, Sm, Sn, Sr, Ta, Tb, Te, Th, Ti, Tl, Tm, U, V, W, Y, Yb, Zn and Zr. All trace elements are presented in parts per million (ppm) and major elements in weight percent (wt %). Pulped samples were digested using a modified aqua regia solution (an equiproportional mixture of HCl, HNO_3 and H_2O) for 1 h in a hot water bath (95 °C). After cooling, the solution was made up to a final volume with 5% HCl and analyzed using an ICP-mass spectrometer (McCurdy et al., 2016). Reliability as assessed by accuracy and precision of the data was determined using field duplicates, analytical duplicates and certified reference materials. Field duplicates were further used to determine the suitability of the data for regional mapping on a per-element basis through an application of Analysis of Variance (ANOVA). The lake sediment geochemical data are intended to capture lithological variation, secondary processes such as alteration and geochemical anomalies related to bedrock mineralization. For one of our workflows, the data was transformed using centered log-ratio (CLR) (Aitchison, 1982). Censored data were replaced with half the detection limit. For a discussion on the appropriateness of the application and effect of log-ratio transformations, particularly as it pertains to data embedding (e.g., feature space geometry), algorithm selection and performance impacts for our intended ML task of geochemical anomaly detection, see Zhang et al. (2021).

3.2. Predictive modelling and mapping

Within the explainable data-driven prospectivity mapping methods, PCA is a common algorithm, since it re-coordinates data along directions of variability, and reduces the dimensionality of chemical coordinates (Carranza, 2008; Zuo, 2011; Grunsky et al., 2014; Harris et al., 2015; Grunsky and de Caritat., 2019). For samples that are mostly composed of minerals, the inter-sample chemical variability is dominated by the variability of mineral composition and chemistry. Hence, the regional geochemical variability can be captured by principal components, which essentially describe key combinations of sample stoichiometries. These multivariate relationships are likely to be reliable for further use, if they are spatially coherent (Grunsky, 2010). Major components may capture regional lithological variability, alteration and mineralization (Grunsky and Smee, 1999; Grunsky, 2010), whereas lesser components likely capture under-sampled or random processes. Eigenvectors are typically manually interpreted (e.g., Grunsky et al., 2014). Thereafter, models can be built for exploration purposes, such as regression residual maps of an element of interest against the components (e.g., Harris et al., 2015; Arne et al., 2018; Grunsky and de Caritat, 2019).

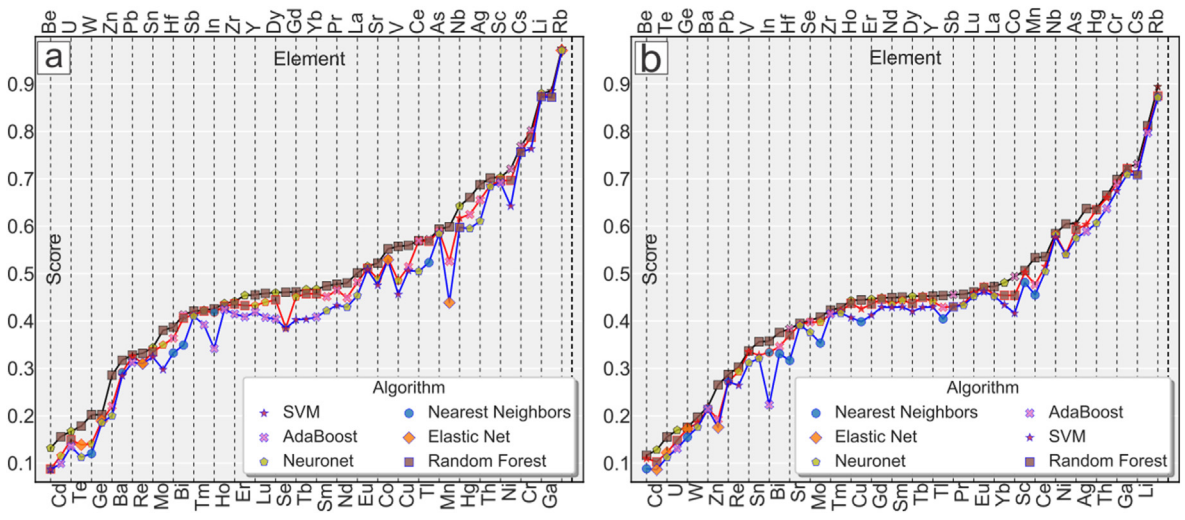

Fig. 4. Results of the algorithm selection and performance assessment using (a) CLR-transformed data, and (b) raw data, as measured using the coefficient of determination (CoD). The top three performing algorithms are shown for each element.

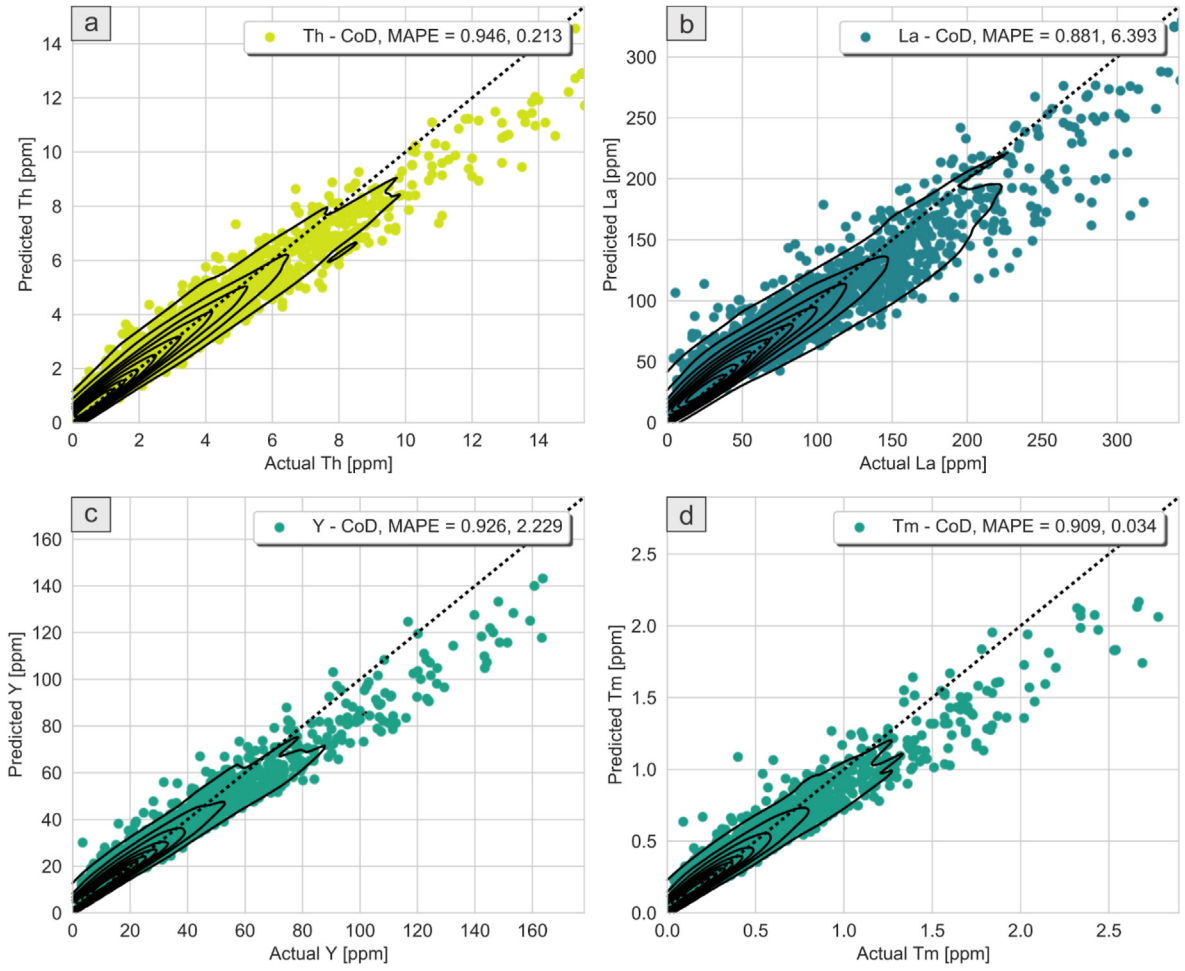

Fig. 5. Scatter plots of prediction results for (a) Th, (b) La, (c) Y and (d) Tm. The best algorithm per element was used. Shown on the graphs are the coefficient of determination (CoD) and median absolute prediction error (MAPE) for each element. Underprediction at high elemental concentrations can be seen.

For our purpose, we showcase the effectiveness of a ML-based technique for prospectivity mapping as developed by Zhang et al. (2021), which makes use of ML to predict trace elemental concentrations using major and minor elemental concentrations. They observed that the trained models effectively serve as geochemical baselines, such that the relative size of geochemical anomalies are proportional to the prediction residuals (predicted minus actual). Maps produced using the prediction residuals demonstrated selectively altered geochemical contrast,

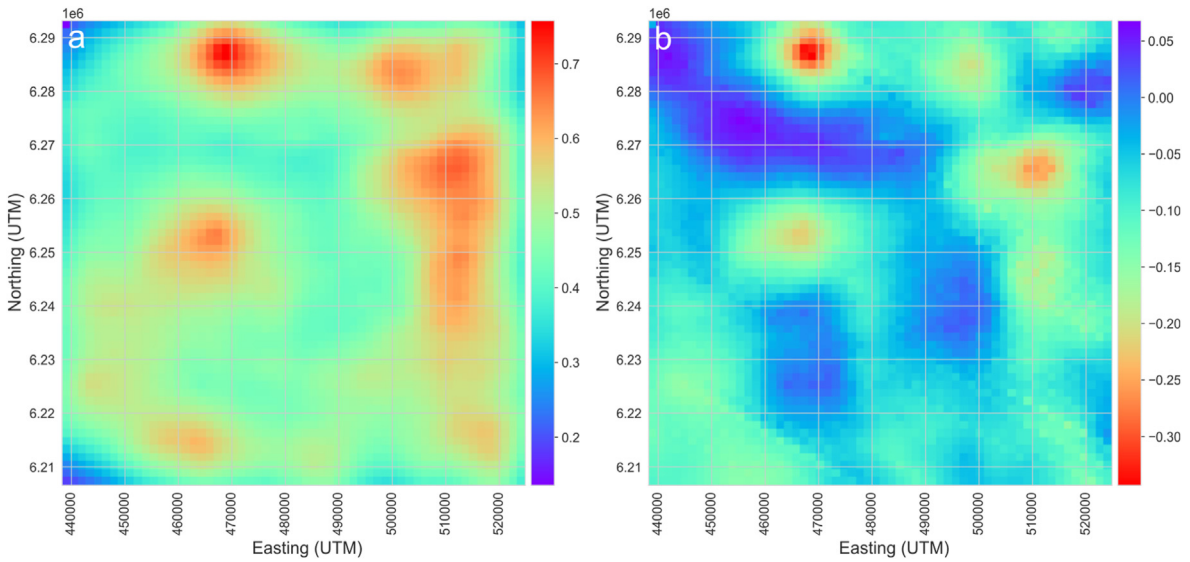

Fig. 6. (a) Stacked elemental concentration maps of most of the chalcophile elements in the dataset, and (b) a corresponding prediction residuals map. Both stacked maps use uniform weights and normalized layers.

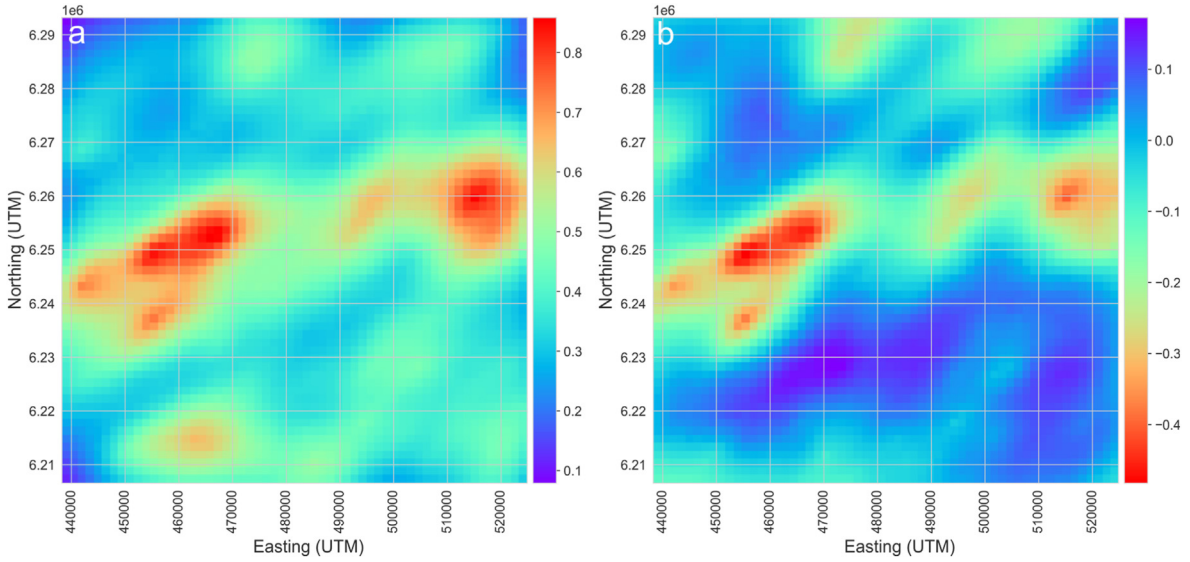

Fig. 7. (a) Stacked elemental concentration maps of most of the REEs in the dataset, and (b) a corresponding prediction residuals map. Both stacked maps use uniform weights and normalized layers.

suppressing large regional variability and increasing contrast between geochemical anomalies and the background (Zhang et al., 2021). A key discipline-specific recognition made by Zhang et al. (2021) was that different elements carried different classes of geoscientific information – major and minor elements carry the rock-forming information of samples, whereas trace elements carry information regarding geological and geochemical processes that may not be related to rock-forming (or even secondary) processes. Therefore, using rock-forming elements as ML features (predictors) and trace elements as targets effectively automates the building of geochemical baselines that are capable of suppression of large-scaled (e.g., lithological) variability in the trace element data. This can be considered a form of discipline-specific feature engineering in the ML context or knowledge application in the prospectivity mapping context. In our observation, this application of discipline-specific knowledge effectively removes the requirement to manually delineate pervasive background processes and targeted anomalies (e.g., using eigenvector interpretation).

Following Zhang et al. (2021), we use major and minor elemental concentrations to predict the concentrations of the trace elements under similar assumptions as Grunsky and de Caritat (2019), in that regional lithology and secondary processes have high explanatory power for a majority of the observed geochemical variability. In this manner, we assume that the key variabilities in lithology, alteration and mineralization are captured by the data of major and minor elemental concentrations. Hence, anomalies that by definition, cannot be explained by the rock-forming elements must constitute either noise or additional geological or geochemical processes, some of which are probably associated with mineralization processes that are of further exploration interest. In our technique, the trained ML models serve as implicit and dynamic (they are capable of predicting trace elemental concentrations across many lithologies or rock types, see Zhang et al., 2021) regional geochemical baselines and the prediction residuals are the deviations in elemental concentrations. Although supervised methods are used, they are strictly used to build geochemical baselines and are not used to train

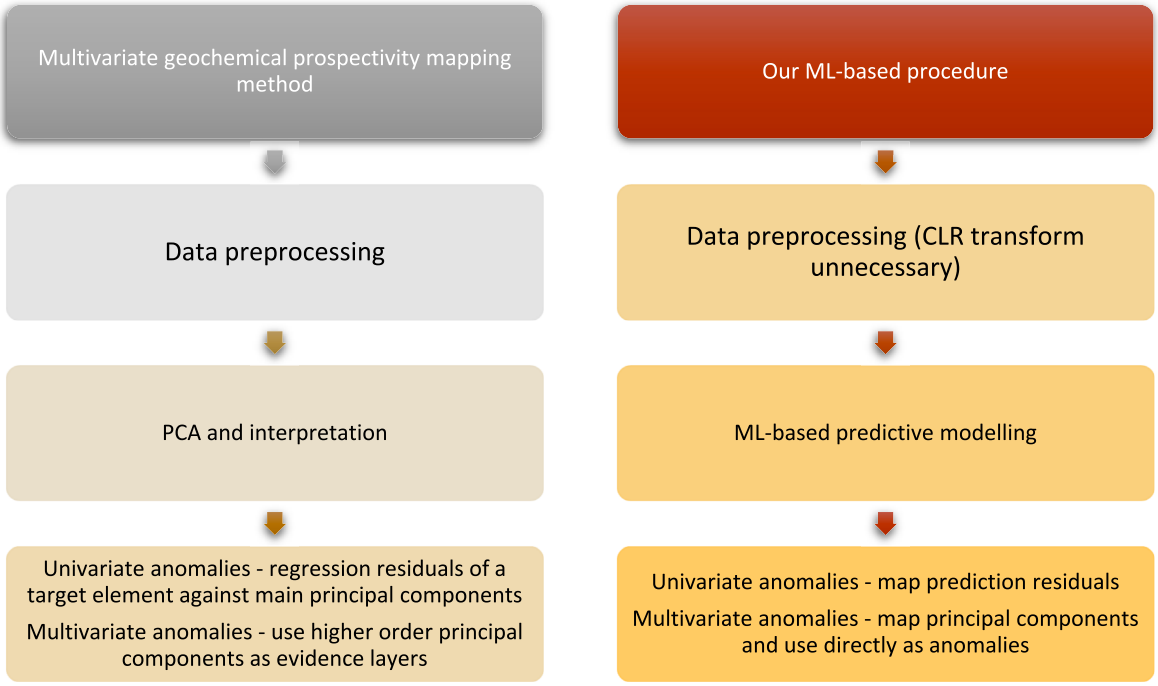

Fig. 8. Workflow comparison for the multivariate geochemical anomaly detection and mapping method versus our procedure.

ML algorithms to recognize anomalies by example. Hence, our anomaly detection method is unsupervised, as anomalies in the area are known to exist (and have been studied extensively) but the data is unlabelled. Against the baselines on a per-element basis, negative prediction residuals (prediction is less than actual) implies that there is an excess concentration above the regional baseline that by definition, cannot be explained by one or more models. The choice of the sign of prediction residuals is arbitrary and depends on how the residuals are calculated (actual minus predicted or predicted minus actual). The quality of the model would be maximized through algorithm selection and parameter tuning, to minimize the prediction residuals using a metric. A more explanatory model better suppresses the large-scale variability in the data compared to less explanatory models. As such, the more explanatory power of the model, the more selective the resulting anomalies in geographical space. Furthermore, the explain-ability of the models can be explicitly balanced against model performance on a per-application basis. This implies that the regional geochemical baseline can be tuned through model selection and model tuning. In our approach, more than one algorithm is used to predict a single element, the best performing algorithm and model can be used for each element (alternatively, a one-algorithm-fits-all approach is feasible to reduce computational requirements). Univariate maps of prediction residuals for each trace element can be produced for direct use as anomaly maps or as intermediate evidence layers for other methods that follow this stage. To produce multivariate anomaly maps, an algorithm such as PCA can be used to summarize the major anomalies in the region. In this case, a multivariate map that is produced from the best performing combination of models is a stacked output of multiple algorithms over multiple elements. Spatially coherent areas of negative prediction residuals are regions of anomalous elemental enrichment, which are worthy of further investigation. Detection of depletion processes are also theoretically possible, although this aspect is unexplored in this paper. Since qualitative ground truth exists in the area (e.g., known targets are REEs and a dispersal trend is well-documented), it is possible to understand the validity of our results.

We created a ML workflow that implements data preprocessing, predictive modelling and visualization. For the purpose of this study, the features were all the major and minor elements (Fe, Mg, Al, Ca, Na, K, Ti, P, S). For our ML task, it is not important whether the features are CLR-transformed or not as part of feature engineering (Hastie et al., 2009; Domingos, 2012), since it is not expected to significantly alter the predictive modelling performance across a range of ML algorithms (studied in more detail in Zhang et al., 2021). However, for one of our workflows, we utilized the CLR transformation to enable a parallel multivariate modelling approach for comparison purposes and also to compare the results between the uses of raw and CLR-transformed ML features. After the CLR transformation, the features were rescaled, such that each feature spans an equal numerical range. During the predictive modelling stage, the features were used to train regression algorithms (Russell and Norvig, 2010). The regression algorithms used in this study include: k-nearest neighbors (Cover and Hart, 1967; Fix and Hodges, 1951; Kotsiantis et al., 2007; Witten and Frank, 2005), random forests (Ho, 1995; Breiman, 1996a, 1996b; Kotsiantis, 2014; Freund and Schapire, 1995; Sagi and Rokach, 2018), elastic net (Santosa and William, 1986; Tibshirani, 1996; Tikhonov, 1943; Zou and Hastie, 2005), support vector machine (SVM; Vapnik, 1998; Hsu and Lin, 2002; Karatzoglou et al., 2006), adaptive boosting of decision trees (AdaBoost; Freund and Schapire, 1995) and artificial neural networks (neuronet; Hastie et al., 2009; Curry, 1944; Lemaréchal, 2012; Rosenblatt, 1961; Rumelhart et al., 1986; Cybenko, 1989). For a detailed description of these algorithms, their assumptions on the feature space properties (e.g., Euclidean geometry and its associated metric) and their parameters, see Zhang et al. (2021).

For model selection and tuning, we employed cross-validation. To detect anomalies in our technique, cross-validation is not strictly required as mathematical induction would not occur and instead a geochemical baseline (a fitted model) is required. Deviations from the baseline can be measured through a metric, e.g., prediction residuals. However, to understand the prediction performance, model generalizability and to prevent over-fitting (exaggeration of the baselines), we use cross-validation to perform algorithm selection and model tuning. We employ the coefficient of determination (CoD) metric to assess prediction accuracy. The choice of the performance metric is important to anomaly detection, as the CoD metric is sensitive to outliers and therefore, optimizing the prediction performance using the CoD metric implies that the geochemical baseline is tuned to suppress outliers and therefore, this optimization suppresses the models' sensitivity to sparsely sampled

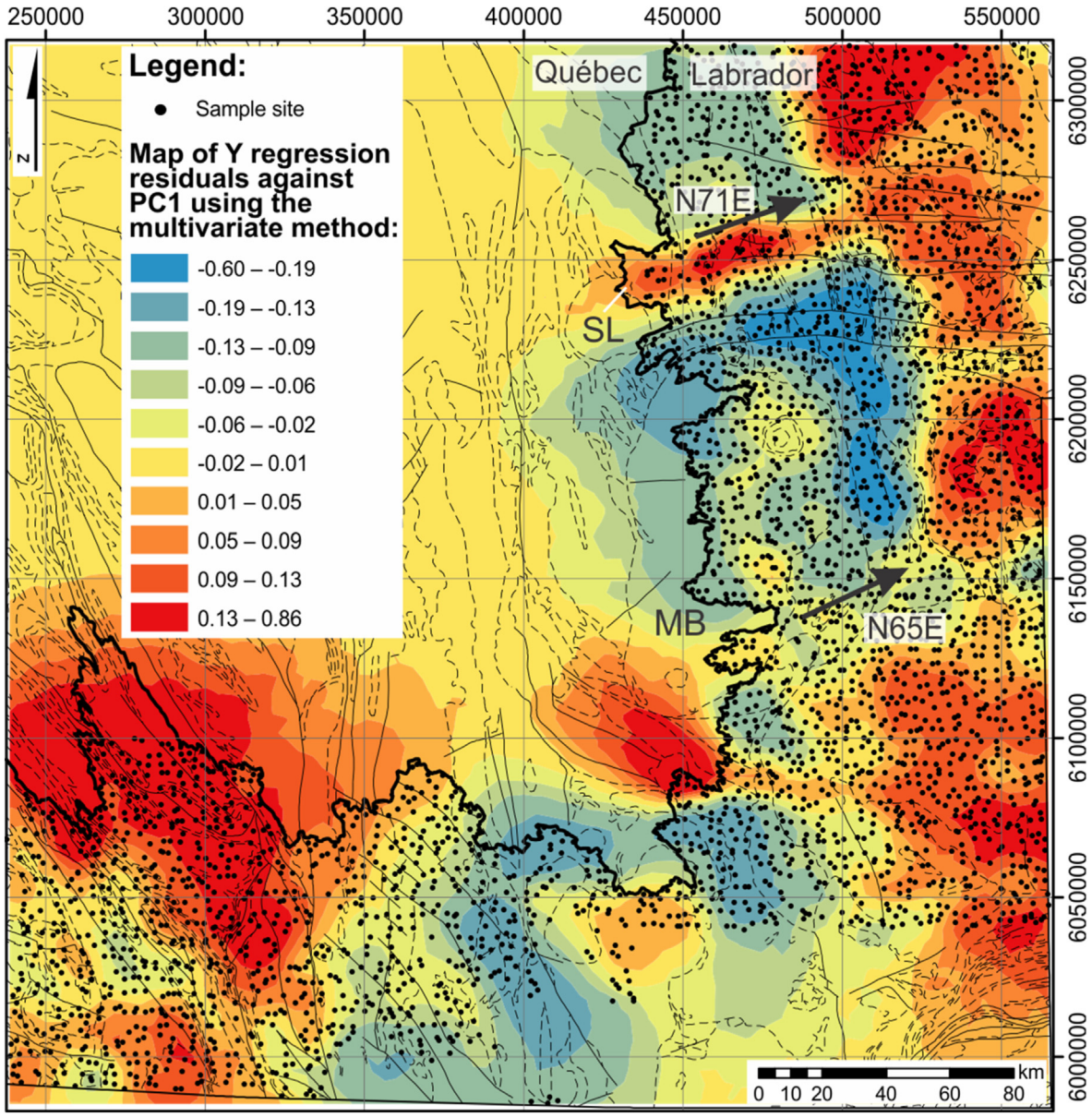

Fig. 9. Regional prospectivity map of Y using the multivariate method. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake peralkaline intrusion is visible and extends approximately 100 km to the northeast.

anomalies. The median absolute prediction error (MAPE) is a measure of median prediction error and here we use it as an additional metric to profile prediction performance but not to perform algorithm selection or model tuning. We used an exhaustive grid search (see Table 1) combined with an n-fold cross validation (n = 4) to determine the rank-order of algorithms. The first 3 top performing algorithms per element are then used to produce prediction results using 10-fold cross-validation over 100 runs. The results were thereafter averaged and the prediction residuals were calculated for each performance level, for each element. Subsequently, the prediction results were used in two ways to generate geochemical anomaly maps, the first uses either weighted evidence layers of prediction residual maps or single element residual maps, which simulates a knowledge-based method to combine evidence layers to target one or more specific indicator elements. The second uses a PCA-based method to generate a linear combination of prediction residuals (across all elements predicted and using a variety of algorithms for a given performance level) to create maps of the most significant multivariate geochemical anomalies.

Our method further advances the data-driven aspect of prospectivity mapping using compositional geochemical data (Grunsky and de Caritat, 2019), as it does not assume underlying stoichiometries nor require principal component interpretation. In addition, our procedure allows for any regression algorithm to be used, and does not assume linearity of elemental relationships, which means that our procedure can incorporate a range of parametric, non-parametric and linear-, nonlinear and neural network-based algorithms and the quality of the mapping would be directly linked to the performance of the algorithms and models, and is not linked to the interpretability of eigenvectors, operator expertise or discipline-specific knowledge. Lastly, we demonstrate that the CLR-transformation is unnecessary for our technique.

4. Results

Prediction results indicate that Ta, Pd, Au, Pt and B are generally

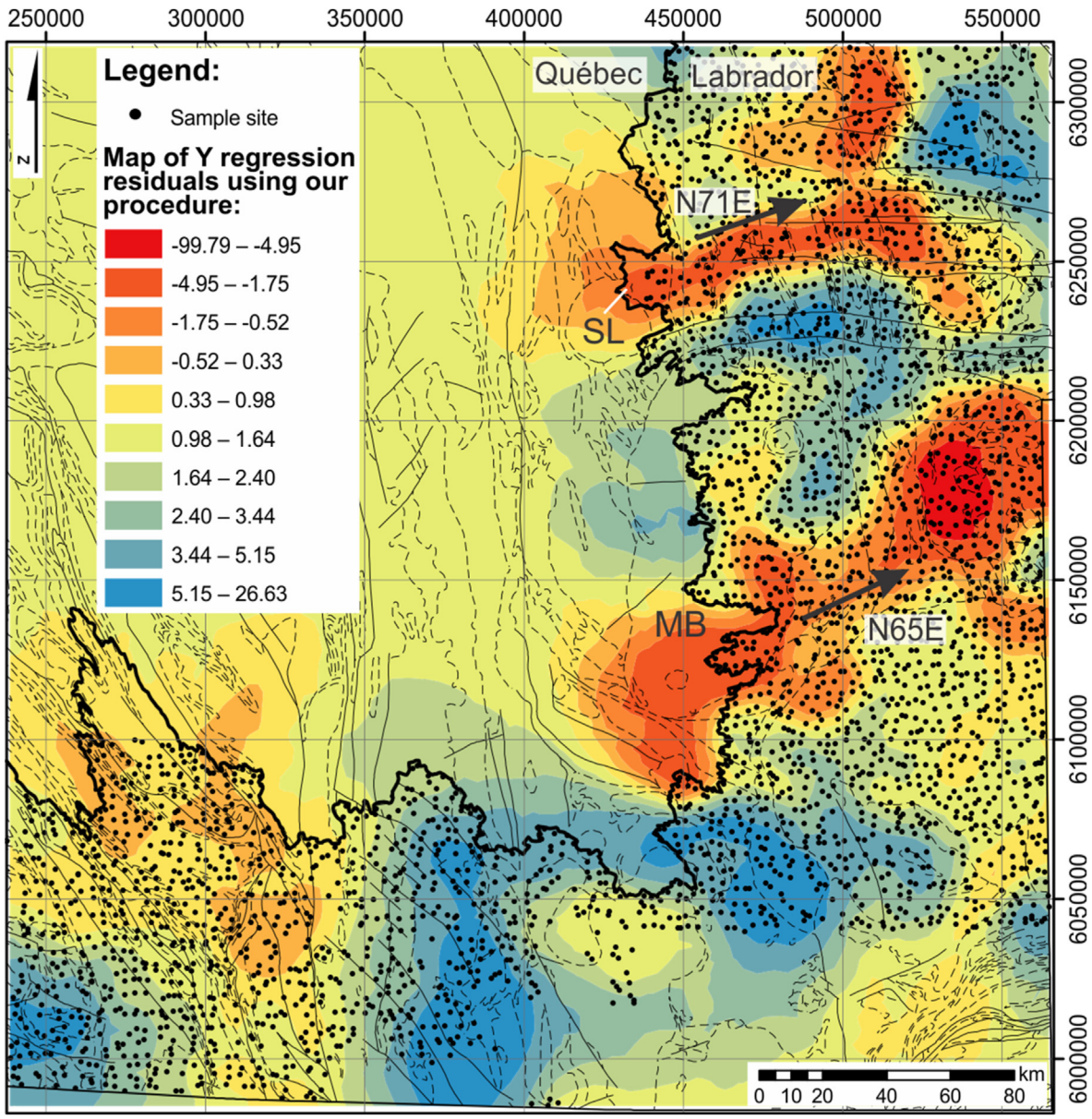

Fig. 10. Regional prospectivity map of Y using our ML-based procedure. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is visible and extends approximately 100 km to the northeast. Furthermore, another glacial dispersal trend is visible extending 140 km northeast from the southern portion of the Mistastin Batholith.

unpredictable (CoD <0.1) due to pervasive data censoring. Hence, they were excluded from further analyses and mapping. Comparison of results between those using the raw versus the CLR-transformed data demonstrate that on average, the raw data produces more accurate predictions (Fig. 3). The best algorithm and its optimal hyperparameters depend on the target element and the type of data used (Fig. 4, see also Table 2). However, within the parameter grid, the random forest regressor is generally the best using either CLR-transformed or raw data (Fig. 4). This is interesting, since the random forest algorithm is not spatially aware in the feature space and therefore, there is no heuristic to suggest any specific embedding geometry. No obvious trends exist for second- and third-best algorithms. However, SVM, Neuronet and AdaBoost tends to perform better than k-nearest neighbors and elastic net, although the performance differential between all algorithms is typically a few percent and sometimes identical to two decimal places. Therefore, complex models can be substituted by simpler and more explainable models with minor losses of anomaly selectivity. The use of the CoD metric implies that the prediction performance is heavily affected by outliers. However, this is desirable if the prediction residuals would be used to detect anomalies, in which case, it is less risky to suppress large anomalies than to generate more false positives. The prediction results show a systematic under-prediction at higher concentrations for REE indicator elements in the region (as identified in McClenaghan et al., 2017), such as Th, La, Tm and Y (Fig. 5). This implies that rock-forming elements are insufficient to explain higher concentrations of REE indicator elements, which likely contain true regional geochemical anomalies, if they are also spatially coherent. Hence, there are likely less well-sampled geological and geochemical processes operating in the region that enriched some samples with REEs relative to the geochemical baseline. Moreover, there is a general consensus between various algorithms and their best models on the relative size of the geochemical anomalies (the prediction residuals are often statistically similar).

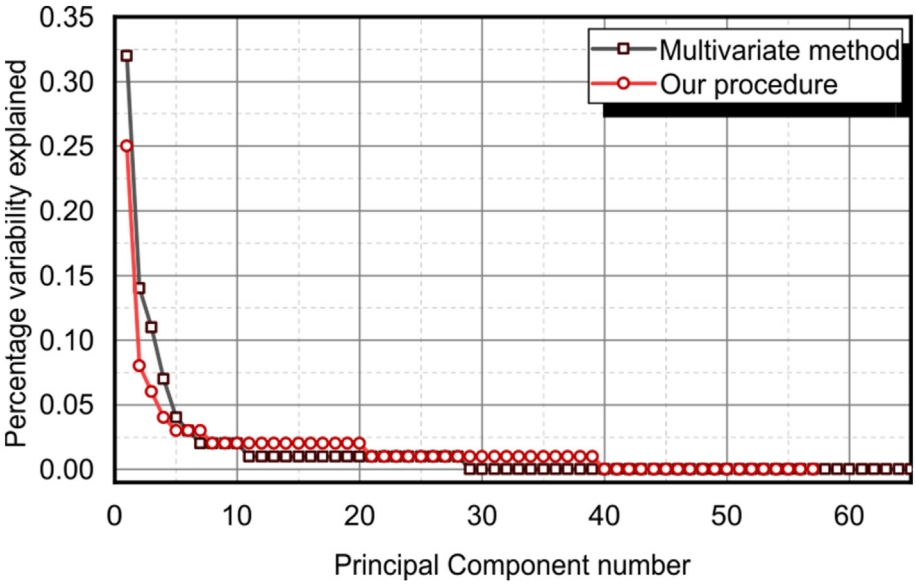

Fig. 11. Scree plot of the principal components using the multivariate prospectivity mapping method and our procedure.

For process validation and to demonstrate an integration of our technique with knowledge-driven methods, we created maps of the original concentrations and the prediction residuals for two groups of elements – REEs (Sc, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb and Lu) and chalcophiles (Cu, Zn, Ga, As, Se, Sn, Sb, Te, Bi, Tl, Pb, In, Hg, Cd, Ge and Ag). For mapping, only elements that exhibited reliable directional semivariograms that can be fitted with spherical models were used. The extent of the maps was chosen to contain the west-to-east glacial dispersal trend (Geological Survey of Canada, 1980; Batterson, 1989; Zajac, 2015; Paulen et al., 2017; McClenaghan et al., 2017, 2019). Element maps were stacked using per-layer normalization and uniform weighting, to give each element an equal influence on the final map. Final maps of the prediction residuals and the elemental concentrations show similar patterns; however, the residual map is less noisy and more selective, which results in higher visible contrast (Figs. 6 and 7). The dispersal trend is not visible for the chalcophile elemental maps, but is visible in the REE-maps (Fig. 7).

There are other methods to create anomaly maps, depending on the nature of the exploration (greenfield or brownfield). In a brownfield setting (or e.g., where there are known target elements), anomalies of indicator or deposit-vectoring elements can be mapped, e.g., univariate prediction residual maps. For greenfield exploration, an additional step is required to automatically extract the most prominent regional geochemical anomalies. For this purpose, many algorithms are available, such as PCA. However, unlike the application of PCA in multivariate geochemical process discovery (e.g., Carranza, 2008; Grunsky and de Caritat, 2019), the principal components of prediction residuals are only expected to contain geochemical anomalies at the scale of the survey. To within model capability, regional variability for all trace elements that can be modelled by multivariate elemental associations using major and minor elements have been removed. Therefore, there is no need to interpret eigenvectors for mapping, beyond an understanding the composition of anomalies. We produced two sets of maps using a multivariate method (e.g., Grunsky and de Caritat, 2019) and our technique. The steps to accomplish both are summarized in Fig. 8.

For a univariate comparison, we chose to compare maps of Y anomalies produced from the multivariate and our methods. For the former method we constructed principal components from CLR-transformed data and verified that the first few principal components capture large-scaled geochemical variation. Subsequently, we used a multilinear regression of the Y concentration against the first few principal components to produce regression residuals that were mapped (Fig. 9). Maps produced using this method, up to principal component 3 are qualitatively identical with the contrast maximized using principal component 1. For our method, the Y prediction residuals were directly mapped (Fig. 10). A comparison of the two maps shows that the map produced using our method captures the glacial dispersal trend emanating from the Strange Lake peralkaline intrusion and corresponds to the pattern previously identified using gamma-ray spectrometry, lake sediments and lake waters. In comparison, the multivariate method yields a Y-enrichment pattern that is similar to the glacial dispersal trend in the west, but connects to a large body to the east. This large body is unobserved using our technique and undocumented in previous studies of the glacial dispersal trend using boulder mapping, matrix till geochemistry, indicator minerals and airborne gamma-ray spectroscopy data (Geological Survey of Canada, 1980; Batterson, 1989; Zajac, 2015; Paulen et al., 2017; McClenaghan et al., 2017, 2019).

For a multivariate comparison, it is impossible to compare maps of principal components of anomalies against those produced by the multivariate method, since these sets of principal components have different meanings. Instead, we make a comparison between the two groups on a basis of roughly the same amount of manual effort (which we deem a fair criterion, as methods that are less data-driven can be arbitrarily refined using better knowledge). For both methods, one or more of the principal components should contain known regional geochemical anomalies – the REE dispersal trend. Hence, we chose maps produced by all methods that exhibited the highest resemblance of the dispersal trend. For our method, the highest performing algorithms produced results that (see Fig. 4) feature a single dominant principal component (principal component 1), which captures approximately 24.74% of the region's anomalies. For the multivariate method, principal component 1 explains 32.42% of the data's variability (Fig. 11). The highest resemblance between the principal component maps occurred for principal components 1 of both methods (with a manual examination of the multivariate method's principal components up to number 6) (Figs. 12 and 13). It is clear that the map of principal component 1, derived from the multivariate method reflects regional geochemical variability (Fig. 12). Principal component-1 map using the multivariate method is insufficient for further exploration targeting and typically, additional procedures, such as a linear discriminant analysis or elemental regressions against a principal component(s), would be used to further refine desirable exploration targets (e.g., desirable lithologies). In comparison, our

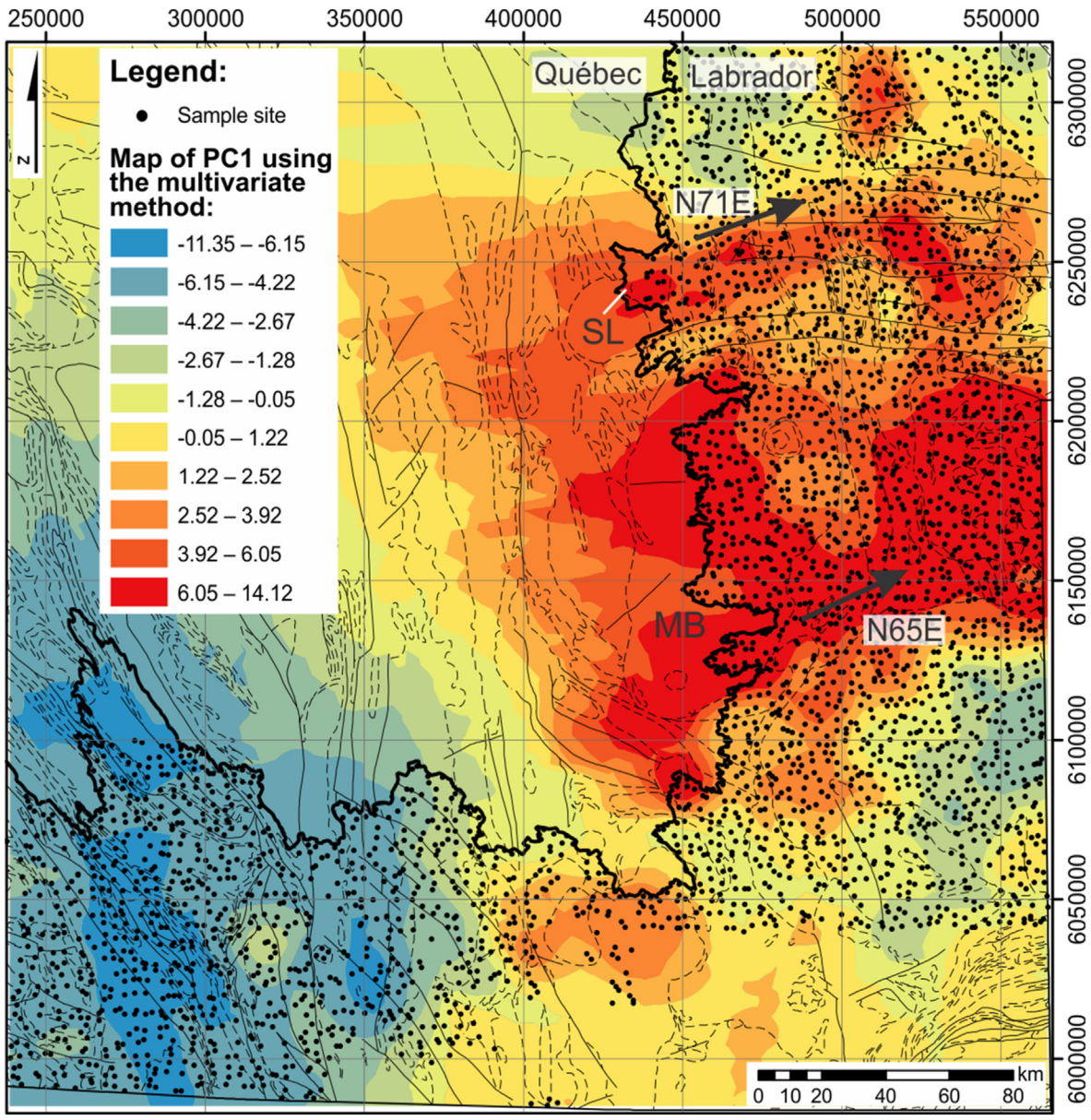

Fig. 12. Map of principal component 1, using the multivariate method. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is poorly distinguishable from the background.

method yields a regional anomaly map that is sufficiently selective to be used for further exploration targeting, especially if combined with glaciation striation directions (Fig. 13).

Anomaly maps that are produced using the raw data, using our method, generally exhibit an increase in overall contrast and particularly in the intensity of the anomalies in the Strange Lake and Mistastin vicinity relative to those produced using the CLR-transformed data, although the differences are unsubstantial (results of the best combinations of algorithms visualized in Figs. 13 and 14). Based on the performance comparison between the two workflows (Fig. 3), this is expected, since the models using the raw data are generally higher performing. There is a trend of a loss of anomaly selectivity (e.g., anomalous regions become larger) with the use of lower performance algorithms (Figs. 13–18) across the three best combinations (as depicted in Fig. 4). The loss in selectivity appears to be somewhat less pronounced for CLR-transformed data using the third-best combination of algorithms, however the contrast remains greater for the map that was derived from the workflow using raw data (Figs. 17 and 18).

5. Discussion

5.1. Geological interpretation and validity of the observed trends

The glacial dispersal trend associated with REE-enriched Strange Lake peralkaline complex is well known and has been extensively documented (Geological Survey of Canada, 1980; Batterson, 1989; Zajac, 2015; Paulen et al., 2017; McClenaghan et al., 2017, 2019). Both univariate prospectivity maps of Y (and similarly for other indicator elements) using the multivariate method (i.e., Grunsky and de Caritat, 2019) and our procedure (Figs. 9 and 10, respectively) clearly demonstrate that a consistent and cohesive pattern is detectable in the lake sediment geochemical data. Lake sediment geochemical data elemental distribution patterns are known to mimic those of glacial till, as indicator minerals and fragments of mineralization found in glacial till inevitably migrate into local drainage systems where they settle/precipitate into lakes (McConnell and Batterson 1987; Cameron, 1994). In prospectivity maps produced using our technique, the dispersal trend can be traced for

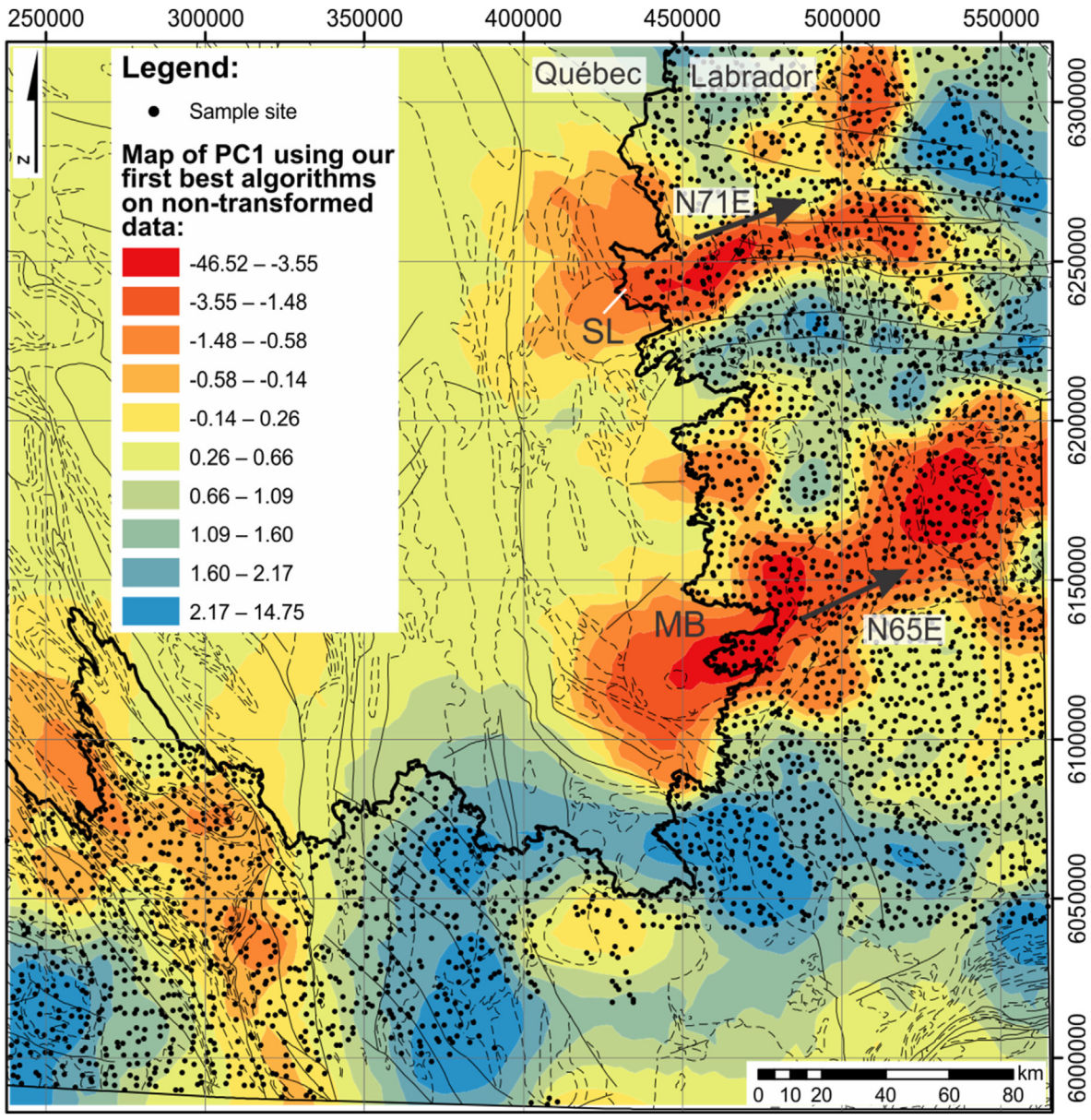

Fig. 13. Map of principal component 1 using the best combination of ML algorithms on raw data, following our procedure. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is visible and extends approximately 100 km to the northeast. Furthermore, another glacial dispersal trend is visible extending 140 km northeast from the southern portion of the Mistastin Batholith.

approximately 100 km to the northeast. This distance is in agreement with that of previous studies. However, compared to the multivariate method, both our univariate and multivariate prospectivity maps clearly highlight another previously undocumented dispersal trend originating from the southern portion of the Mistastin Batholith (Figs. 10 and 13). Recent mapping and exploration of the southern portions of the Mistastin Batholith in 2010 by Quest Rare Minerals and Midland Exploration have found a number REE-bearing deposits that also contain Zr, Y, U, Th, Be, Pb and Nb (Hammouche et al., 2012). Glacial erosion features in the southern portions of the Mistastin Batholith strike N65E, which corresponds to the glacial dispersion trend emanating from the batholith. It therefore appears that the dispersal train originating from the southern portion of the Mistastin Batholith is also detectable in the lake sediment geochemical data. However, the composition of the dispersal trends is different between the southern and the northern dispersal trends. A full interpretation of the region's geochemical anomalies is out of the scope of this paper. Lastly, another less-prominent anomaly can be seen southwest of the Mistastin Batholith (Figs. 10 and 13). A geological investigation in the area does not reveal any known deposits or lithologies that could be associated with REE, Zr, Y, ± Be, Nb, U, Th deposits. Furthermore, the anomaly is not associated with the local glacial dispersal trend (towards the southeast). Therefore, further work in the area would be required to identify the nature of this anomaly.

5.2. *Anomaly detection-based prospectivity mapping*

There have been many approaches to using geochemical compositional data, much of it within the multivariate domain. For the purpose of prospectivity mapping, the key goal is to produce maps that correctly identify, with as high selectivity as possible, true geochemical anomalies that are invariably indicative of the locations of mineral occurrences. Although the definition of geochemical anomaly has not changed, the methods used to delineate baselines from anomalies have continued to evolve and improve. There has been a general trend of further adoption

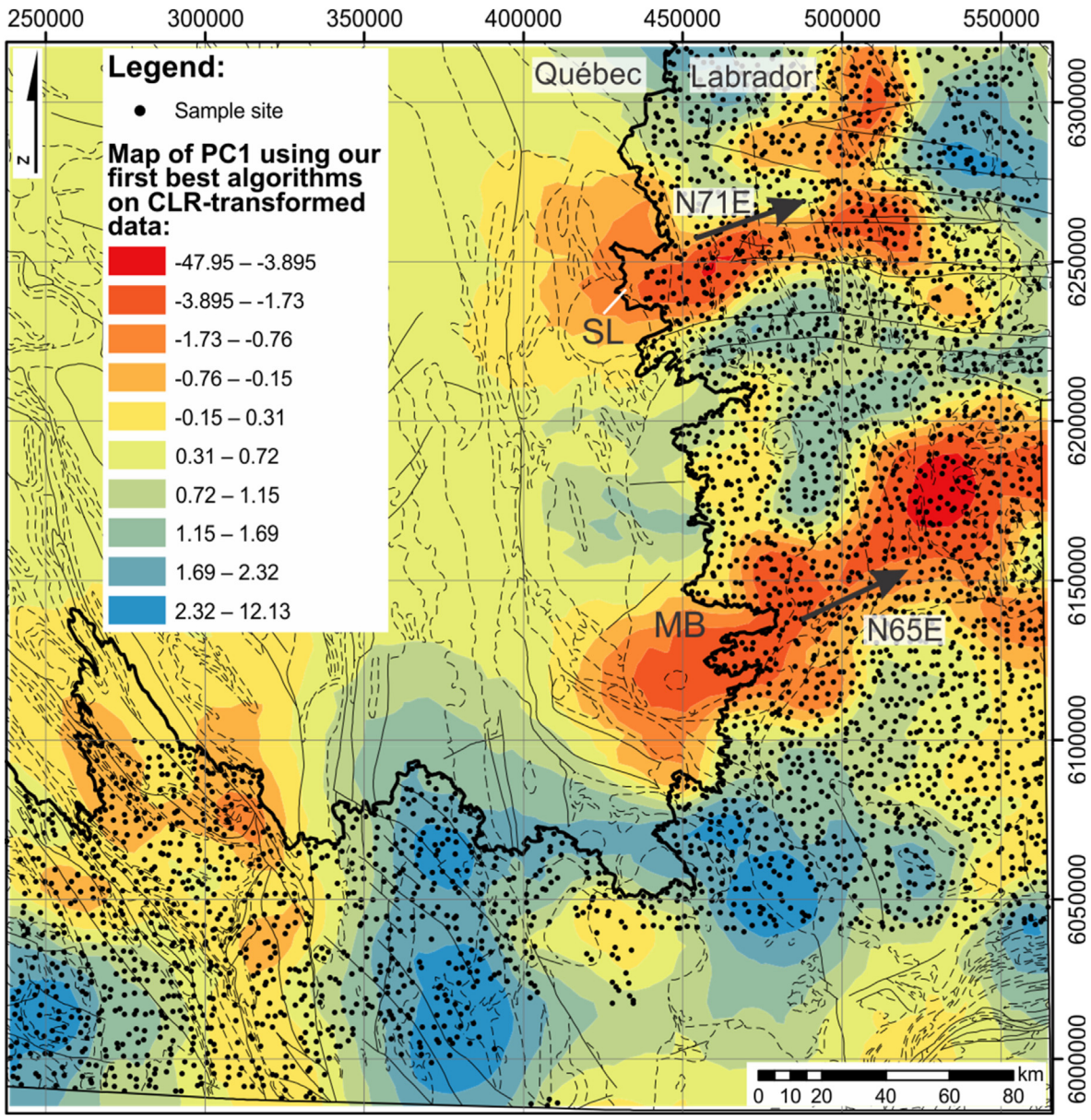

Fig. 14. Map of principal component 1 using the best combination of ML algorithms on CLR-transformed data, following our procedure. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is visible and extends approximately 100 km to the northeast. Furthermore, another glacial dispersal trend is visible extending 140 km northeast from the southern portion of the Mistastin Batholith.

of data-driven techniques, which can complement or replace knowledge-driven techniques, and especially in brownfield exploration, because the existence of at least partial ground truth allows for model validation. A general change in the discipline of geochemical exploration has been towards larger regional surveys, higher quality data and more capable and sometimes more complex data modelling methods. Consistent with this trend, in this study, we build upon the multivariate modelling approach as demonstrated in Grunsky and de Caritat (2019) by generalizing the idea of geochemical anomaly detection, automating much of the data analytics and modelling using a modern ML-based pipeline and removing the intermediary steps that require traditional geological and geochemical interpretation (i.e., of the principal components). Consequently, as a product of our approach:

(a) we do not require log-ratio transformations for geochemical anomaly detection, although they facilitate a parallel workflow that uses multivariate modelling methods for comparison;
(b) we do not assume the existence of minerals or stoichiometry in the data, which implies that our technique is more broadly applicable and does not require interpretation of stoichiometries (as demonstrated using volcanic rock geochemical data in Zhang et al., 2021);
(c) we are able to directly map prediction residuals of any combination of elements to produce prospectivity maps that are more selective and less noisy than that of the multivariate approach;
(d) we are able to produce multivariate prospectivity maps that do not require extensive manual interpretation and the resulting maps are substantially more selective and reflective of the 'ground truth';
(e) we do not require log-ratio feature transformations and therefore remove the necessity for data imputation.

A key functional difference in our approach compared to the multivariate method is that we utilize ML extensively and designed our

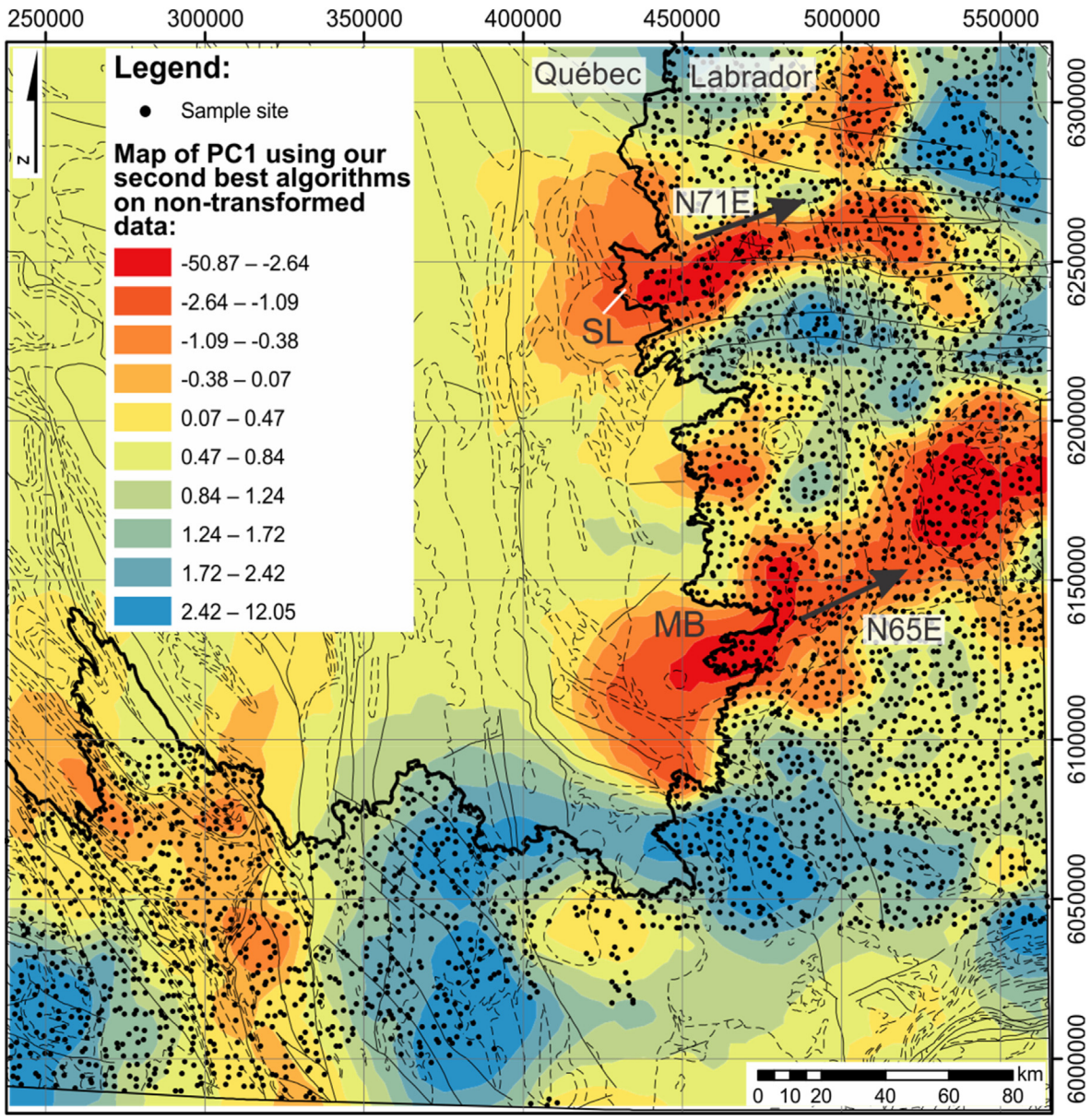

Fig. 15. Map of principal component 1 using the second-best combination of ML algorithms on raw data, following our procedure. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is visible and extends approximately 100 km to the northeast. Furthermore, another glacial dispersal trend is visible extending 140 km northeast from the southern portion of the Mistastin Batholith.

workflow to be agnostic of specific algorithms. The function of the ML algorithms as a data pre-processor for multivariate modelling is to automate the construction of a regional geochemical baseline, against which, all anomalies are measured. The choice of algorithm and model are entirely automated, and the best algorithm and model parameters are built without manual interpretation of regional geology (e.g., lithology or alteration) or knowledge of the type of mineral deposit being sought. In addition, ground truth is not required to train the models (as the anomaly detection method is unsupervised). This means that the quality of the geochemical baseline, and therefore the value of the detected geochemical anomalies is dependent on the quality of the predictive modelling. Models with high explanatory powers (e.g., a high CoD) translate to reliable regional geochemical baselines. As such, a weakness of our approach lies in that most of the data should capture non-anomalous regional geochemistry. Since the baselines and anomalies are determined through the data using ML algorithms, the definition of anomaly is not a knowledge-driven one and changes depending on the data used to train the models. This is unavoidable and delegates the interpretation of the nature of the anomalies (what elements are in them) to the users of the prospectivity maps, where anomalous compositions would have to be understood regardless of the method of discovery. Our results show that a single principal component of the prediction residuals immediately produces a map that qualitatively agrees with the regional ground truth (the Strange Lake peralkaline intrusion REE glacial dispersal trend). Compared to a similar application of PCA in the multivariate method, we do not require additional principal components to be mapped, interpreted, or further manipulated, at least for this dataset. This is because the ML algorithm training that occurred during the predictive modelling stage already accounted for typical regional geochemical variability in the trace elements, and as such, a single dominant principal component captures the most important regional geochemical anomalies (compare Figs. 12 and 13). Therefore, our method is expected to be more objective and produce more consistent and replicable prospectivity maps than the multivariate method, as

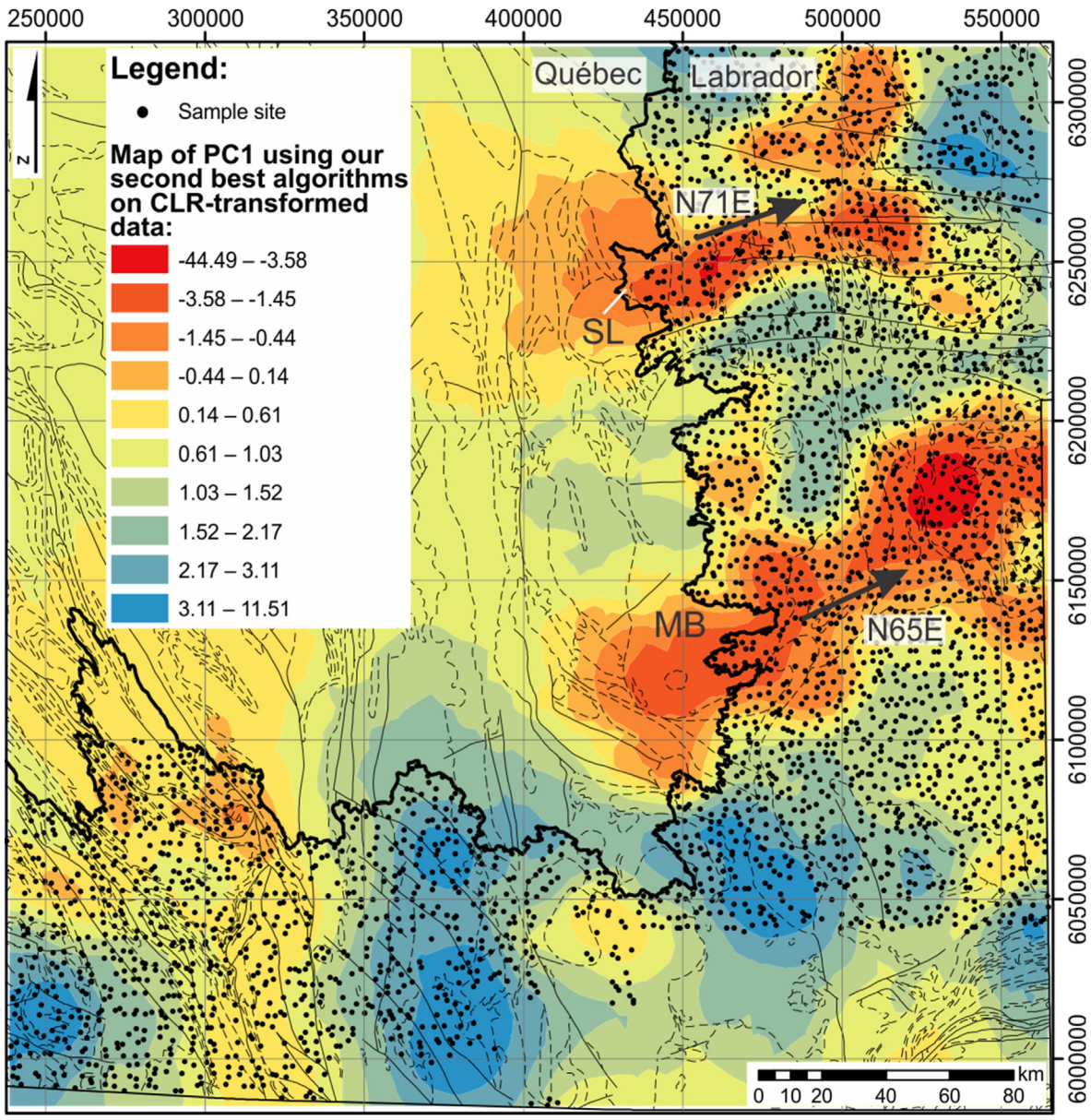

Fig. 16. Map of principal component 1 using the second-best combination of ML algorithms on CLR-transformed data, following our machine procedure. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is visible and extends approximately 100 km to the northeast. Furthermore, another glacial dispersal trend is visible extending 140 km northeast from the southern portion of the Mistastin Batholith.

manual interpretation is not utilized in intermediary steps. When our approach is used in the univariate application to generate elemental anomaly maps, it is clear that our approach is superior to the multivariate method, which requires a convoluted approach to first identify key stoichiometries in the area that vary, then regress a target element against the main principal components to determine the regression residuals, which are used to produce anomaly-type of prospectivity maps. In addition, we show that we do not require log-ratio data transformations and show that if data transformations such as log-ratio transformations are to be applied to a workflow, then their effects should be quantitatively measured against a baseline (e.g., using raw data). Where it can be shown that they improve the predictive modelling performance, then we anticipate that the application of those transformations can similarly result in a higher contrast in resulting geochemical anomaly maps. For these types of comparisons, performance resulting from a workflow using raw data should be used as a baseline or default, since any transformation must be demonstrably performance-enhancing. In our approach, since we do not assume the existence of stoichiometry in the data, nor do we require extensive manual interpretation, our method is more data driven and more flexible. However, as our approach is data driven, the quality of baselines depends on the quality of the data, in the same sense as for multivariate modelling, in that the data needs to capture a substantial geochemical baseline. This is not a drawback that is unique to our approach, but is general to all data-driven approaches, since anomalies can only be contextualized against a large backdrop of non-anomalous data. Lastly, our method is designed to be able to accommodate a range of ML algorithms, such that we leave the choice between complex and powerful, and simple and highly explainable algorithms to the user through an informed knowledge of their performance differences.

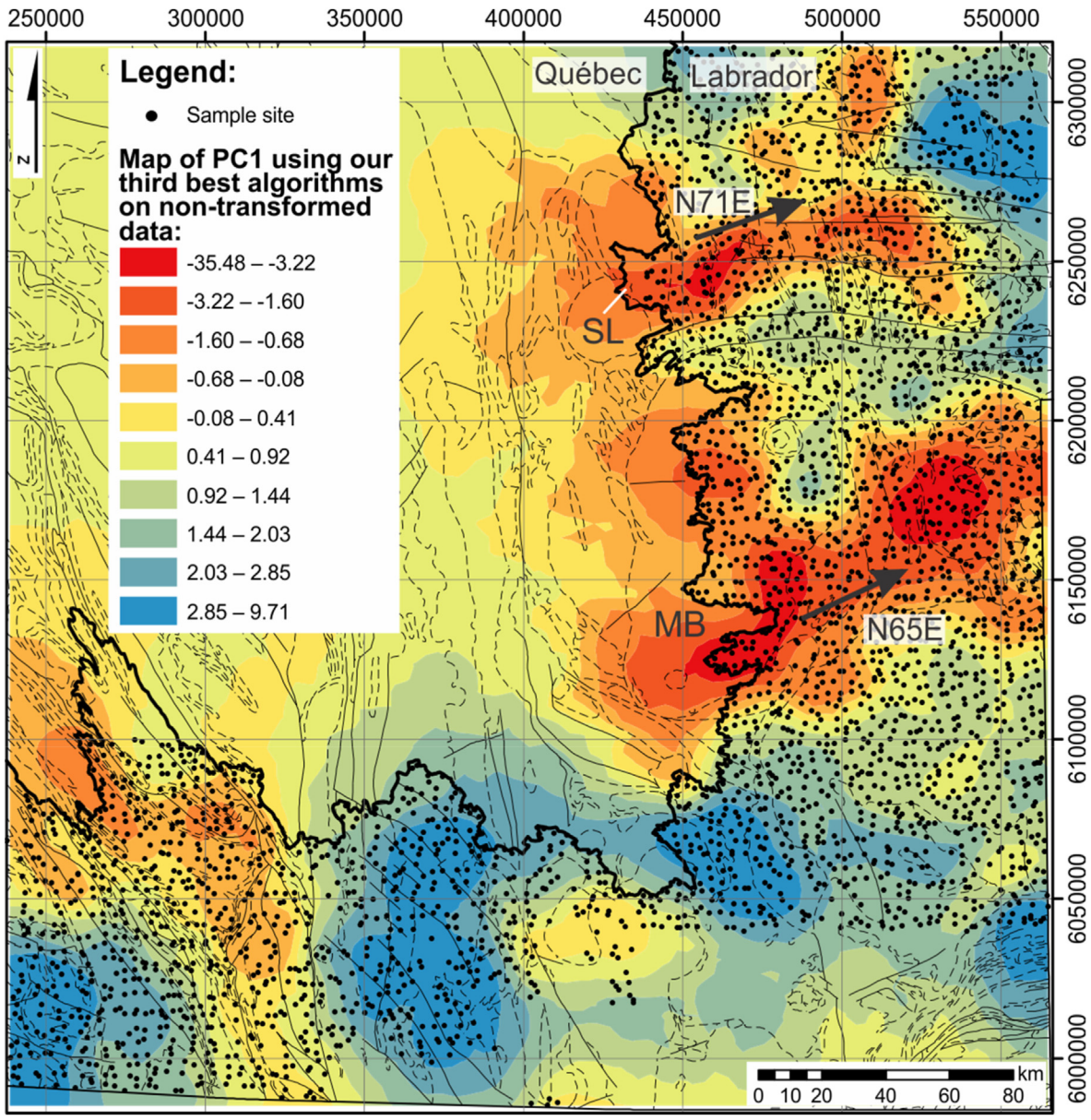

Fig. 17. Map of principal component 1 using the third-best combination of ML algorithms on raw data, following our procedure. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is visible and extends approximately 100 km to the northeast. Furthermore, another glacial dispersal trend is visible extending 140 km northeast from the southern portion of the Mistastin Batholith.

6. Conclusion

The use of geochemical data continues to evolve and prospectivity mapping has always greatly benefitted by developments in data analysis, modelling and mapping methods. With the proliferation of data-driven methods such as ML, it is possible to simultaneously improve upon existing approaches, while generating entirely new approaches. In this study, we demonstrated a technique to leverage compositional data by using ML methods to improve upon the existing multivariate approach and as a substitute for it in some applications. In comparison with the multivariate method, our technique does not assume stoichiometric variability of elements, does not require geoscientific knowledge-based process discovery and interpretation, does not require data imputation and is flexible with respect to the choice of algorithm to balance the need for performance with the need for model explain-ability. Therefore, our approach is more data-driven and the workflow itself is almost entirely automatable, which requires little manual interaction. This facilitates the adoption of our technique, especially where the amount of data or projects exceeds the amount of manual labour. For its application in lake sediment geochemistry and for the exploration of REEs in the southeast Churchill Province, our procedure is both more accurate and more selective than the multivariate method, to within a comparable amount of manual labour. We expect that our approach would generalize well to other types of exploration geochemical data.

Declaration of conflicting interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

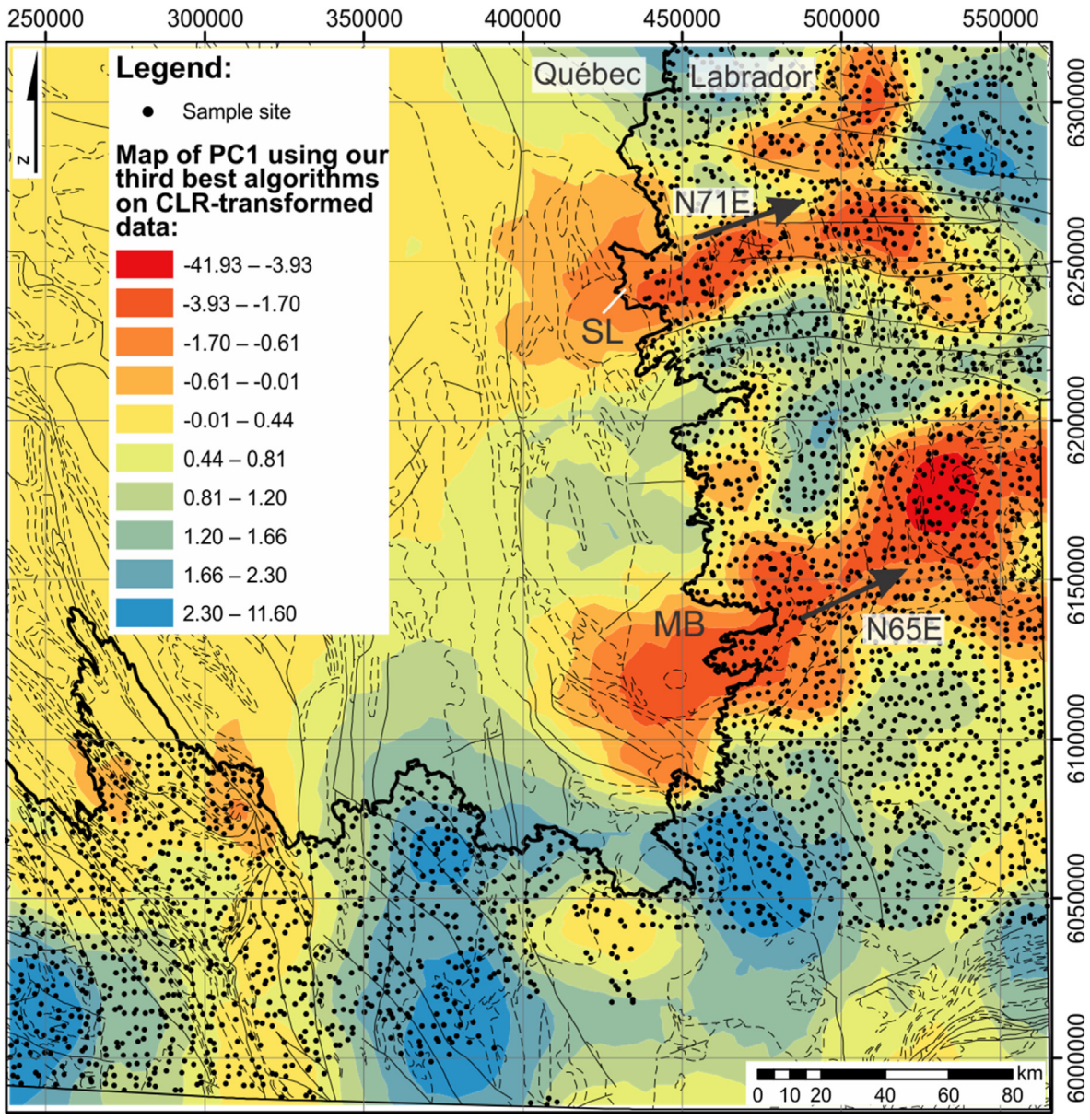

Fig. 18. Map of principal component 1 using the third-best combination of ML algorithms on CLR-transformed data, following our procedure. Included in the regional prospectivity map are the lithological boundaries (Fig. 2) and glacial dispersion trends (black arrows). Abbreviations: Strange Lake peralkaline intrusion (SL); Mistastin Batholith (MB). The glacial dispersal trend from the REE-bearing Strange Lake intrusion is visible and extends approximately 100 km to the northeast. Furthermore, another glacial dispersal trend is visible extending 140 km northeast from the southern portion of the Mistastin Batholith.

Acknowledgements

The authors would like to thank Annick Morin (GSC) for providing geological mapping support for the study area. We would also like to thank Beth McClenaghan (GSC) and Dr. Robert F. Garrett (GSC, Emeritus scientist) and three anonymous reviewers for providing insightful comments that have greatly improved the readability and scope of this manuscript. The authors thank Dr. Hua Wang for editorial handling. Data used in this study is open access and can also be found at the Natural Resources Canada (NRCan) publication database (GEOSCAN) under OF 8026 or ID 298834.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aiig.2022.02.002. This dataset is published under the Open Government License (https://open.canada.ca/en/open-government-licence-canada).

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Series B. 44 (2), 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

Arne, D., Mackie, R., Pennimpede, C., 2018. Catchment analysis of re-analyzed regional stream sediment geochemical data from the Yukon. Explore 179, 1–30.

Balaram, V., 2019. REEs: 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.

Batterson, M.J., 1989. Glacial dispersal from the Strange Lake alkalic complex, northern Labrador. In: DiLabio, R.N.W., Coker, W.B. (Eds.), Drift Prospecting. Geological Survey of Canada, pp. 31–40. https://doi.org/10.4095/127362. Paper 89-20.

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists: Modeling with GIS. Pergamon, Elsevier Science, New York.

Breiman, L., 1996a. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

Breiman, L., 1996b. Stacked regressions. Mach. Learn. 24 (1), 49–64. https://doi.org/10.1007/BF00117832.
Brown, W.M., Gedeon, T.D., Groves, D.L., Barnes, R.G., 2000. Artificial neural networks; a new method for mineral prospectivity mapping. Aust. J. Earth Sci. 47 (4), 757–770. https://doi.org/10.1046/j.1440-0952.2000.00807.x.
Cameron, I., 1994. Lake sediment sampling in mineral exploration. In: Govett, G.J.S. (Ed.), Handbook of Exploration Geochemistry. Elsevier, pp. 227–267. https://doi.org/10.1016/B978-0-444-81854-6.50013-4.
Carranza, E.J.M., 2008. Geochemical Anomaly and Mineral Prospectivity Mapping in GIS. Elsevier.
Carranza, E.J.M., Hale, M., Faassen, C., 2008. Selection of coherent deposit-type locations and their application in data-driven mineral prospectivity mapping. Ore Geol. Rev. 33, 536–558. https://doi.org/10.1016/j.oregeorev.2007.07.001.
Chen, S., Hattori, K., Grunsky, E.C., 2018. Identification of sandstones above blind uranium deposits using multivariate statistical assessment of compositional data, Athabasca Basin, Canada. J. Geochem. Explor. 188, 229–239. https://doi.org/10.1016/j.gexplo.2018.01.026.
Chung, C.F., Agterberg, F.P., 1980. Regression models for estimating mineral resources from geological map data. Math. Geol. 12, 473–488. https://doi.org/10.1007/BF01028881.
Corrigan, D., Wodicka, N., McFarlane, C., Lafrance, I., Rooyen, D.V., Bandyayera, D., Bilodeau, C., 2018. Lithotectonic framework of the Core zone, southeastern Churchill province, Canada. Geosci. Can. 45 (1), 1–24. https://doi.org/10.12789/geocanj.2018.45.128.
Cover, T., Hart, P., 1967. Nearest Neighbor Pattern Classification, 13. IEEE Trans. Inf. Theory, pp. 21–27. https://doi.org/10.1109/TIT.1967.1053964.
Curry, H.B., 1944. The method of steepest descent for non-linear minimization problems. Q. Appl. Math. 2, 258–261. https://doi.org/10.1090/qam/10667.
Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Math. Control. Signals, Syst. 2 (4), 303–314. https://doi.org/10.1007/BF02551274.
David, J., Simard, M., Bandyayera, D., Goutier, J., Hammouche, H., Pilote, P., Leclerc, F., Dion, C., 2012. Datations U-Pb effectuées dans les provinces du Supérieur et de Churchill en 2010-2011. Ministère des Ressources Naturelles. Québec, RP, p. 33, 2012-01.
European Commission, 2011. Tackling the challenges in commodity markets and on raw materials. Communication from the commission to the European parliament, Europ. eco. soc. comm. comm. reg. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0025:FIN:en:PDF. (Accessed 5 January 2021).
Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87. https://doi.org/10.1145/2347736.2347755.
European Commission, 2014. On the review of the list of critical raw materials for the EU and the implementation of the raw materials initiative. Communication from the commission to the European parliament, the council. the Europ. eco. soc. comm. comm. reg. Retrieved January 5, 2021, from https://ec.europa.eu/growth/sectors/raw-materials/specific-interest/critical_en.
European Commission, 2017a. On the 2017 list of Critical Raw Materials for the EU, 2017. Communication from the commission to the European parliament, the council. the Europ. eco. soc. comm. comm. reg. https://ec.europa.eu/transparency/regdoc/rep/1/2017/EN/COM-2017-490-F1-EN-MAIN-PART-1.PDF. (Accessed 5 January 2021).
European Commission, 2017b. Study on the review of the list of critical raw materials, criticality assessments. Retrieved January 5, 2021, from. https://op.europa.eu/en/publication-detail/-/publication/08fdab5f-9766-11e7-b92d-01aa75ed71a1.
European Commission, 2020. Critical raw materials resilience: charting a path towards greater security and sustainability. Communication from the commission to the European parliament. Europ. eco. soc. comm. comm. reg. Retrieved January 5, 2021, from https://ec.europa.eu/docsroom/documents/42849.
Fix, E., Hodges, J.L., 1951. An important contribution to nonparametric discriminant analysis and density estimation. Int. Stat. Ins. 57, 233–238. https://doi.org/10.2307/1403796.
Freund, Y., Schapire, R.E., 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P. (Ed.), Second European Conference on Computational Learning Theory. Sprigner.
Friske, P.W.B., McCurdy, M.W., Day, S.J.A., 1996a. National Geochemical Reconnaissance, Labrador Compilation: Distribution of Fluoride in 18,839 Lake Sediment Samples and 1,162 Stream Sediment Samples, Newfoundland (Labrador). Geological Survey of Canada. https://doi.org/10.4095/208315. Open File 3260d.
Friske, P.W.B., McCurdy, M.W., Day, S.J.A., 1996b. National Geochemical Reconnaissance, Labrador Compilation: Distribution of Lead in 18,839 Lake Sediment Samples and 1,162 Stream Sediment Samples, Newfoundland (Labrador). Geological Survey of Canada. https://doi.org/10.4095/208317. Open File 3260f.
GEM, 2019. Evaluation of the Geo-Mapping for Energy and Minerals (GEM-2) Program. Government of Canada, Natural Resources Canada. https://www.nrcan.gc.ca/transparency/reporting-and-accountability/plans-and-performance-reports/strategic-evaluation-division/reports-and-plans-year/evaluation-the-geo-mapping-for-energy-and-minerals-gem-2-program/evaluation. (Accessed 25 August 2021).
Geological Survey of Canada, 1980. Airborne Gamma Ray Spectrometric Map, Dihourse Lake Quebec–Newfoundland. Geological Survey of Canada, Geophysical Series. https://doi.org/10.4095/124881. Map 35124(08)G.
Ghorbani, Y., 2018. Gaining access to high-tech and critical raw materials: a driving factor in developing future concept and methods in mineral processing and extractive metallurgy. In: III Congress of Industrial Engineering of the North of Chile, 22-24th August, 2018, Antofagasta-Chile.
Gowans, R.M., Lewis, W.J., Shoemaker, S., Spooner, J., Zalnieriunas, R.V., 2014. NI-43-101 Technical Report on the Preliminary Economic Assessment (PEA) for the Strange Lake Property, Quebec, Canada. Micron International Limited for Quest Rare Minerals Ltd, p. 258.
Grunsky, E.C., Mueller, U.A., Corrigan, D., 2014. A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: application for predictive geological mapping. J. Geochem. Explor. 141, 15–41. https://doi.org/10.1016/j.gexplo.2013.07.013.
Grunsky, E.C., de Caritat, P., 2019. State-of-the-art analysis of geochemical data for mineral exploration. Geochem. Explor. Environ. Anal. 20, 217–232. https://doi.org/10.1144/geochem2019-031.
Grunsky, E.C., 2010. The interpretation of geochemical survey data. Geochem. Explor. Environ. Anal. 10, 27–74. https://doi.org/10.1144/1467-7873/09-210.
Grunsky, E.C., Arne, D., 2020. Mineral-resource prediction using advanced data analytics and ML of the QUEST-South stream-sediment geochemical data, Southwestern British Columbia, Canada. Geochem. Explor. Environ. Anal. 23 (1). https://doi.org/10.1144/geochem2020-054 geochem2020-g2054.
Hammouche, H., Legouix, C., Goutier, J., Dion, C., 2012. Géologie de la région du lac Zeni. Ministère des Ressources Naturelles. Québec, RG 2012-02, p. 35.
Harris, D.A., Pan, R., 1999. Mineral favourability mapping: a comparison of artificial networks, logistic regression and discriminant analysis. Nat. Resour. Res. 8, 93–109. https://doi.org/10.1023/A:1021886501912.
Harris, J.R., Grunsky, E., Behnia, P., Corrigan, D., 2015. Data-and knowledge-driven mineral prospectivity maps for Canada's North. Ore Geol. Rev. 71, 788–803. https://doi.org/10.1016/j.oregeorev.2015.01.004.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123. Santiago.
Ho, T.K., 1995. Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montréal, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994. Canada.
Hoffman, P.F., 1988. United plates of America, the birth of a craton: early Proterozoic assembly and growth of Laurentia. Annu. Rev. Earth Planet. Sci. Lett. 16, 543–603. https://doi.org/10.1146/annurev.ea.16.050188.002551.
Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Network. 13, 415–425. https://doi.org/10.1109/72.991427.
James, D.T., Connelly, J.N., Wasteneys, H.A., Kilfoil, G.J., 1996. Paleoproterozoic lithotectonic divisions of the southeastern Churchill Province, western Labrador. Can. J. Earth Sci. 33, 216–230. https://doi.org/10.1139/e96-019.
James, D.T., Dunning, G.R., 2000. U–Pb geochronological constraints for paleoproterozoic evolution of the Core zone, southeastern Churchill province, northeastern laurentia. Precambrian Res. 103, 31–54. https://doi.org/10.1016/S0301-9268(00)00074-7.
Karatzoglou, A., Meyer, D., Hornik, K., 2006. Support Vector Machines in R. J. Stat. Softw, pp. 1–28. https://doi.org/10.18637/jss.v015.i09, 15.
Kotsiantis, S.B., 2014. Integrating global and local application of naive bayes classifier. Int. Arab J. Inf. Technol. 11 (3), 300–307.
Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised ML: a review of classification techniques. Emerg. artif. intell. appl. comp. eng. 160 (1), 3–24.
Lemaréchal, C., 2012. Cauchy and the gradient method. Doc. Math. Extra 251 (254), 10.
Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2021. Explainable AI: a review of ML interpretability methods. Entropy 23 (1), 1–18. https://doi.org/10.3390/e23010018.
Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. 29, 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002.
Luo, Z., Xiong, Y., Zuo, R., 2020. Recognition of geochemical anomalies using a deep variational autoencoder network. Appl. Geochem. 122, 104710. https://doi.org/10.1016/j.apgeochem.2020.104710.
McClenaghan, M.B., Paulen, R.C., Kjarsgaard, I.M., Averill, S.A., Fortin, R., 2017. Indicator Mineral Signature of the Strange Lake REE Deposit, Quebec and Newfoundland and Labrador. Geological Survey of Canada, p. 50. https://doi.org/10.4095/306239. Open File 8240.
McClenaghan, M.B., Paulen, R.C., Kjarsgaard, I.M., 2019. Rare metal indicator minerals in bedrock and till at the Strange Lake peralkaline complex, Quebec and Labrador, Canada. Can. J. Earth Sci. 56 (8), 857–869. https://doi.org/10.1139/cjes-2018-0299.
McConnell, J.W., Batterson, M.J., 1987. The Strange Lake Zr-Y-Nb-Be-REE deposit, Labrador: a geochemical profile in till, lake and stream sediment, and water. J. Geochem. Explor. 29, 105–127. https://doi.org/10.1016/0375-6742(87)90073-2.
McCurdy, M.W., Amor, S.D., Finch, C., 2016. Regional Lake Sediment and Water Geochemical Data, Western and Central Labrador (NTS 13-L, 13-M, 14-D, 23-I and 23-J). Geological Survey of Canada, p. 14. https://doi.org/10.4095/298834. Open File 8026.
Miller, R.R., 1990. The Strange Lake pegmatite–aplite-hosted rare-metal deposit, Labrador. In: Current Research. Newfoundland Department of Mines and Energy, pp. 171–182. Report 90-1.
Moorhead, J., Hynes, A., 1990. Nappes in the internal zone of the Labrador Trough: evidence for major, early NW-vergent basement transport. Geosci. Can. 17 (4), 241–244.
Natural Resources Canada, 2021. Critical Minerals. Government of Canada. Retrieved. https://www.nrcan.gc.ca/our-natural-resources/minerals-mining/critical-minerals/23414. (Accessed 24 August 2021).
Overland, I., 2019. The geopolitics of renewable energy: debunking four emerging myths. Energy Res. Social Sci. 49, 36–40. https://doi.org/10.1016/j.erss.2018.10.018.

Paulen, R.C., Stokes, C.R., Fortin, R., Rice, J.M., Dubé-Loubert, H., McClenaghan, M.B., 2017. Dispersal trains produced by ice streams: an example from Strange Lake, Labrador, Canada. In: Tschirhart, V., Thomas, M.D. (Eds.), Proceedings of Exploration '17: Sixth Decennial International Conference on Mineral Exploration, pp. 871–875. Toronto.

Perreault, S., Hynes, A., 1990. Tectonic evolution of the kuujjuaq terrane, new Québec orogen. Geosci. Can. 17 (4), 238–240.

Rizzo, A., Goel, S., Grilli, M.L., Iglesias, R., Jaworska, L., Lapkovskis, V., Novak, P., Postolnyi, B.O., Valerini, D., 2020. The Critical Raw Materials in cutting tools for machining applications. A rev. Mat. 13 (6), 1377. https://doi.org/10.3390/ma13061377.

Rosenblatt, F., 1961. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC. https://doi.org/10.1007/978-3-642-70911-1_20.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L., PDP research group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundation, ume 1. MIT Press.

Russell, S., Norvig, P., 2010. Artificial Intelligence: A Modern Approach, 3rd. Prentice-Hall, New Jersey, pp. 1–1151.

Sagi, O., Rokach, L., 2018. Ensemble learning: a survey. Wiley Interdiscip. Rev.-Data Min. Knowl. Discov. 8 (4), e1249. https://doi.org/10.1002/widm.1249.

Santosa, F., William, W.S., 1986. Linear inversion of band-limited reflection seismograms. J. Sci. Stat. Comput. 7, 1307–1330. https://doi.org/10.1137/0907087.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tikhonov, A.N., 1943. On the stability of inverse problems. Dokl. Akad. Nauk SSSR 39, 195–198.

U.S. Geological Survey, 2018. Draft critical mineral list – summary of methodology and background information. In: U.S. Geological Survey Technical Input Document in Response to Secretarial Order No. 3359. Retrieved January 5, 2021, from. https://pubs.usgs.gov/of/2018/1021/ofr20181021.pdf.

Van der Leeden, J., 1995. Géologie de la région du lac Mistinibi, Territoire du Nouveau-Québec. Ministère des Ressources naturelles. Québec, Rep. MB 95–45, 107.

Vapnik, V., 1998. Statistical Learning Theory. Springer, New York.

Wardle, R.J., James, D.T., Scott, D.J., Hall, J., 2002. The southeastern Churchill Province: synthesis of a Paleoproterozoic transpressional orogeny. Can. J. Earth Sci. 39, 639–663. https://doi.org/10.1139/e02-004.

Witten, I.H., Frank, E., 2005. Data Mining: Practical ML Tools and Techniques, second ed. Morgan Kaufman, San Francisco.

Wright, D.F., Bonham-Carter, G.F., 1996. VHMS favourability mapping with GIS-based integration models, chisel lake–anderson lake area. In: Bonham-Carter, G.F., Galley, A.G., Hall, G.E.M. (Eds.), EXTECH I: A Multidisciplinary Approach to Massive Sulphide Research in the Rusty Lake–Snow Lake Greenstone Belts, Manitoba, 426. Geological Survey of Canada, Bulletin, pp. 339–376.

Zajac, I.S., 2015. John jambor's contributions to the mineralogy of the Strange Lake peralkaline complex, quebec–labrador, Canada. Can. Mineral. 53, 885–894. https://doi.org/10.3749/canmin.1400051.

Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Ashwal, L.D., 2021. Machine-learning-based prediction of trace element concentrations using data from the Karoo large igneous province and its application in prospectivity mapping. Artif. Intell. Geosci. https://doi.org/10.1016/j.aiig.2021.11.002 accepted.

Zou, H., Hastie, T., 2005. Regularisation and variable selection via the elastic net. J. R. Stat. Soc. Series B 67 (2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Zuo, R., 2011. Identifying geochemical anomalies associated with Cu and Pb–Zn skarn mineralization using principal component analysis and spectrum–area fractal modeling in the Gangdese Belt, Tibet (China). J. Geochem. Explor. 111 (1–2), 13–22. https://doi.org/10.1016/j.gexplo.2011.06.012.

Zuo, R., Xiong, Y., Wang, J., Carranza, E.J.M., 2019. Deep learning and its application in geochemical mapping. Earth Sci. Rev. 192, 1–14. https://doi.org/10.1016/j.earscirev.2019.02.023.