Artificial Intelligence in Geosciences 3 (2022) 86–100

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

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

Advanced geochemical exploration knowledge using machine learning: Prediction of unknown elemental concentrations and operational prioritization of Re-analysis campaigns

Steven E. Zhang [a,b], Julie E. Bourdeau [a,b], Glen T. Nwaila [b,*], Yousef Ghorbani [c]

[a] Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, K1A 0E8, Canada
[b] Wits Mining Institute, University of the Witwatersrand, 1 Jan Smuts Ave., Johannesburg, 2000, South Africa
[c] Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, SE–971 87, Luleå, Sweden

ARTICLE INFO

Keywords:
Machine learning
Re-purposing legacy data
Advanced exploration knowledge
Canada
Geochemical data

ABSTRACT

In exploration geochemistry, advances in the detection limit, breadth of elements analyze-able, accuracy and precision of analytical instruments have motivated the re-analysis of legacy samples to improve confidence in geochemical data and gain more insights into potentially mineralized areas. While a re-analysis campaign in a geochemical exploration program modernizes legacy geochemical data by providing more trustworthy and higher-dimensional geochemical data, especially where modern data is considerably different than legacy data, it is an expensive exercise. The risk associated with modernizing such legacy data lies within its uncertainty in return (e.g., the possibility of new discoveries, in primarily greenfield settings). Without any advanced knowledge of yet unanalyzed elements, the importance of re-analyses remains ambiguous. To address this uncertainty, we apply machine learning to multivariate geochemical data from different regions in Canada (i.e., the Churchill Province and the Trans-Hudson Orogen) in order to use legacy geochemical data to predict modern and higher dimensional multi-elemental concentrations ahead of planned re-analyses. Our study demonstrates that legacy and modern geochemical data can be repurposed to predict yet unanalyzed elements that will be realized from re-analyses and in a manner that significantly reduces the latency to downstream usage of modern geochemical data (e.g., prospectivity mapping). Findings from this study serve as a pillar of a framework for exploration geologists to predictively explore and prioritize potentially mineralized districts for further prospects in a timely manner before employing more invasive and expensive techniques.

1. Introduction

Geochemical exploration necessitates the methodical collection, processing and analysis of collected samples. In large exploration programs (regional or national, decadal programs), this process takes time on the order of years to decades to map significant portions of target areas. Coincidentally, the evolution of analytical instruments has been roughly decadal from research to development, to deployment en masse (Balaram, 2021). This overlapping of timescales of advances in analytical instrumentation and the operationalization of large exploration programs unfortunately means that, in a cyclical fashion, geochemical data becomes outdated and resampling and/or re-analyses are required. Outdated geochemical data is data that is analytically no longer the state-of-the-art (e.g., less analyzed elements than modern data). We hereon refer to outdated geochemical data as 'legacy' geochemical data, although aside from specific changes in the analytical methodology, such data is identical to current geochemical data in terms of the requirement of manual labour during data generation. Within the data generation portion of geochemical exploration, the methodology experienced drastic changes mainly in the analytical aspects relative to the sampling and processing methods, as analytical instruments have improved substantially in accuracy and precision, detection limits and the breadth of analyzable elements (Cohen et al., 2010; Balaram, 2021). At the moment, there is still further on-going research to adopt new techniques, such as Raman spectroscopy; and a move towards instrument miniaturization (Balaram, 2021). The cyclical modernization of geochemical exploration data is necessary in some cases, for example, to add additional elements that was absent in legacy data. However,

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

https://doi.org/10.1016/j.aiig.2022.10.003
Received 19 May 2022; Received in revised form 30 October 2022; Accepted 31 October 2022
Available online 5 November 2022

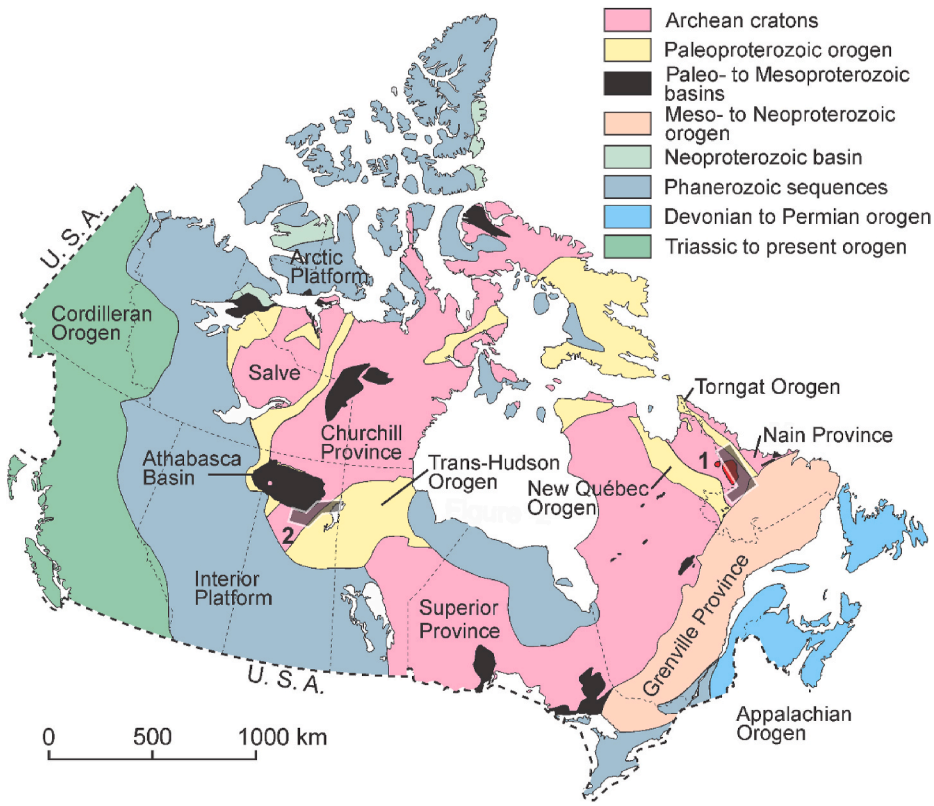

Fig. 1. Simplified geological map of Canada with the location of the lake sediment datasets used in this study. The dataset corresponding to the southeastern Churchill Province (shaded area labelled 1) is located in Labrador. Included in the southeastern Churchill Province is the location of the Mistatin Batholith (in red). The dataset corresponding to the southwestern Churchill Province and Trans-Hudson Orogen (shaded area labelled 2) is located in Saskatchewan.

accomplishing this task using solely manual re-sampling and/or re-analysis is costly and maybe untimely for some elements. However, predictive data generation could be used to guide modernization of geochemical exploration data, and for some uses, even substitute for actual primary data.

Large regional and national geochemical survey programs are a key source of geochemical data. For example, the Geo-mapping for Energy and Minerals (GEM) program (Lebel, 2020; Geological Survey of Canada, 2022) is a Canadian national survey program that was initially conceived in 2008 to provide modern, publicly available geoscience knowledge of northern Canada to support increased exploration for resources (now known as the GEM-GeoNorth program [2020–2027]). Similar programs exist worldwide. For example, the Council for Geoscience in South Africa conducted a series of geochemical surveys that lasted into the early 2000s and are now expanding the survey using modern methods that better cover trace elements and particularly energy-transition and battery metals (Council for Geosciences, 2022). Viewing the GEM program as a prototype, the downstream purpose of the data resulting from these programs seems to be: construction of digital maps; provision of information that may be relevant to assess mineralization potential; and provision of information necessary to make informed decisions on general land use (GEM-GeoNorth, 2022). Depending on the natural landscape of the region being surveyed and the capability of the surveyor, many types of sample media could be available, such as: bulk rock, lake sediment, glacial till and water. Each type of media has its strengths and weaknesses, collection and processing methodology. A primary decision criterion in choosing sample media is its abundance and accessibility within a target area. For every type of sample media, in general, the resulting data (including legacy data) is produced by applying a highly rigorous and standardized methodology from collection to instrumental analysis. This greatly aids data usage, including data fusion.

In general within geochemical sampling, there exists a practice to sample more material than strictly necessary, in anticipation of sample reuse and therefore avoiding re-sampling, in part motivated by perpetual improvements in analytical capability. Re-analysis campaigns are necessary to provide state-of-the-art (modern) data and result in a high and sustained operational cost to geochemical exploration programs, especially at the national level. However, since samples are unchanged between subsequent analyses (where sufficient remnant material exists), they provide an unprecedented opportunity to leverage machine learning to predict elemental concentrations ahead of re-analysis campaigns to aid campaign planning through target prioritization. This enables better resource allocation and planning, which is even more crucial where exploration programs have particular and immediate focus directions – e.g., for the discovery of deposits containing critical raw materials. For this application, where legacy data do not contain elements belonging to a list of critical raw materials, machine learning, legacy geochemical data and modern geochemical data can be used to predict into unknown areas, or areas of interest that were previously targeted for exploration of mineral deposits. In addition to prioritization, this can aid in the early detection of potentially relevant chemical signatures associated with mineral deposits, even before elements belonging to the chemical signatures are physically analyzed.

This proof-of-concept study uses lake sediment data from the GEM program. Lake sediment geochemical data has been generated for decades to explore surficial geology in Canada, because of its good spatial coverage. Re-analyses have to date, produced two fully modernized datasets in 2016 and 2021 respectively: a portion of the southeastern Churchill Province in Labrador (McCurdy et al., 2016); and a portion of the southwestern Churchill Province and Trans-Hudson Orogen in Saskatchewan (Bourdeau et al., 2021, 2022). Initial sampling was conducted in 1978 and 1982 (southeastern Churchill Province), and 1984 and 1986 (southwestern Churchill Province and Trans-Hudson Orogen).

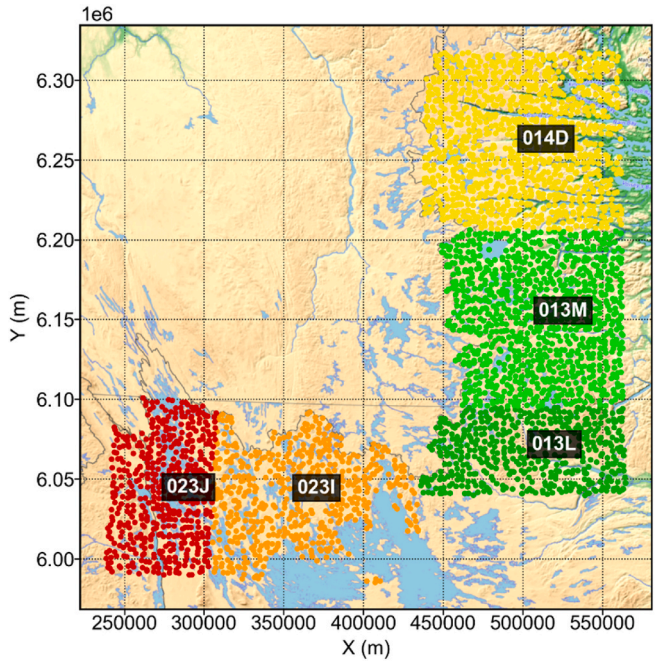

Fig. 2. Sample locations from the southeastern Churchill Province dataset, located in Labrador.

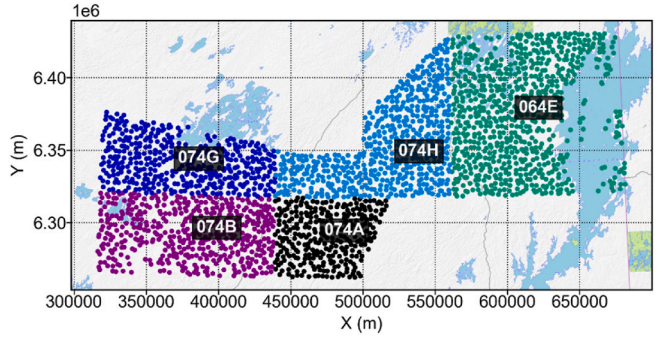

Fig. 3. Sample locations from the southwestern Churchill Province and Trans-Hudson Orogen dataset, located in northeastern Saskatchewan.

As part of a re-analysis campaign in the 1990s, the southeastern Churchill Province in Labrador area was also re-analyzed. In this fashion, the GEM program intends to re-analyze its collection of lake sediment samples to modernize its surficial geochemical data. However, GEM does not currently make use of predictive prioritization in its program execution in any capacity and targets that may be of interest to a modern economy (e.g., mineral deposits containing critical raw materials) are best prioritized on a partial basis of a-priori knowledge of the geology of target areas (knowledge-driven). To the best of our knowledge, there is no geochemical exploration program that is currently performing predictive data generation.

Aside from differences in scientific data quality, a primary structural difference between the legacy and modern geochemical datasets (of sediments or otherwise) lies in the former's lack of a breadth of elements. For example, between 1970 and 1980s, lake sediment geochemical analysis at the Geological Survey of Canada data routinely contained up to 14 elements and did not include any of the rare earth elements. Similarly, for geochemical data that was generated by the Council for Geoscience in the 2000s, a common configuration is 23 elements (e.g., van Rooyen et al., 2004). This makes it impossible to leverage such legacy geochemical data to understand the national distribution of unanalyzed trace elements, many of which have become interesting, such as critical raw materials. In comparison, the newest lake sediment geochemical datasets from GEM now contain 65 elements, ranging from lithophiles, siderophiles and chalcophiles (e.g., McCurdy et al., 2016). Analysis and downstream modelling for exploration using, for example, elemental maps, elemental ratios, multivariate and artificial intelligence approaches are all limited by the sparsity of elements available in legacy geochemical datasets. If probable data containing concentrations of currently unavailable elements were available in an area, it could be leveraged to perform many tasks, such as creating relatively simple elemental maps. This is a task suitable for predictive

Table 1
Details for geochemical datasets used in this study.

OPEN REPORT NUMBER	YEAR	NTS AREAS	NUMBER OF SAMPLES	RE-ANALYSIS	ELEMENTS ANALYZED
557	1978	013L, M	1,425		Ag, As, Co, Cu, F,
559	1978	014D	1,001		Fe, Hg, LOI, Mn,
560	1978	023I, J	1,239		Mo, Ni, Pb, U, Zn
904	1982	023I, J	430		
1359	1986	074-A, B, G and H	1,286		Ag, As, Au, Cd, Co, Cu, Fe, Hg, LOI, Mn, Mn, Mo, Ni, Pb, Sb, U, V, Zn
1643	1984	064-E, 074A and H	1,177		Ag, As, Au, Ba, Br, Cd, Ce, Co, Cr, Cs, Eu, Fe, Hf, Ir, La, Lu, Mo, Na, Ni, Rb, Sb, Sc, Se, Sm, Sn, Ta, Tb, Te, Th, U, W, Yb, Zn, Zr
8026	2016	014D, 013M and L, 023I and J	3,441	557, 559, 560, 904	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, Zr
8837	2021	074-A, B, G and H	1,286	1359	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, Zr
8871	2022	064-E, 074A and H	1,177	1643	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, Zr

Table 2
Comparison of precision between older (1984–1986) and newer (2021) geochemical data for the southwestern Churchill Province and Trans-Hudson Orogen, located in northeastern Saskatchewan. LDL = lower detection limit, and RSD = relative standard deviation. All elemental means, unless otherwise stated, are given in ppm. The method to calculate precision is detailed in McCurdy and Garrett (2016).

Element	1984–1986 data via AAS			2020–2021 data via ICP-MS		
	Elemental mean	Percentage censored LDL	RSD (%)	Elemental mean	Percentage censored LDL	RSD (%)
Ag	0.24	95.42	49.22	0.01	0	13.62
As	8.39	58.17	4.08	4.71	21.97	14.15
Au		100		0.00	37.88	67.67
Cd	0.37	43.79	26.42	0.51	0	6.38
Co	7.28	2.61	11.65	8.13	0	3.56
Cu	14.31	0	8.00	14.06	0	3.62
Fe (%)	5.15	0	7.48	5.17	0	4.29
Hg	0.01	0	11.05	0.01	0	14.11
Mn	911.42	0	21.22	480.56	1.52	6.89
Mo	4.09	55.56	21.39	3.24	0	4.40
Ni	13.52	0	6.10	17.36	0	4.55
Pb	7.53	90.2	6.85	3.31	0	7.66
Sb	0.25	94.12	9.43	0.08	8.33	41.33
U	8.24	1.31	8.65	7.74	0	4.27
V	38.24	2.61	9.96	37.45	0	4.45
Zn	100.73	0	7.72	101.53	0	5.14

Table 3
Random forest parameter grid.

PARAMETER	RANGE
MAXIMUM DEPTH	5, 7, 9, 11, unlimited
MAXIMUM FEATURES	2, 3, 4, 5, 6, 7, 8, unlimited
MINIMUM SAMPLES PER SPLIT	3, 4, 5, 6, unlimited
MINIMUM SAMPLES PER LEAF	1, 2, 3, 4, 5, 6

modelling using machine learning. In the past, various machine learning algorithms have been demonstrated to be able to predict elemental concentrations (Zhang et al., 2021a, 2022). In particular, for our purpose, machine learning algorithms can be used to model the relationships between elements for matching samples in old and new geochemical datasets. The models can then be deployed to infer probable concentrations of yet unanalyzed elements in other areas ahead of their actual re-analyses.

In this study, we examine the feasibility of predicting elemental concentrations in areas with legacy data ahead of their re-analysis using machine learning, legacy and modern geochemical datasets. We show that a range of elements can be predicted with a range of performance profiles (measured using metrics). In addition, the spatial concordance between the predicted and actual elemental concentrations are generally substantially better than that in the variable domain and should be sufficient, such that predicted elemental concentration maps provide advanced exploration knowledge that could be used for operational prioritization and other downstream uses. We expect that our approach would be generalizable to other environments and exploration campaigns, provided that legacy and modern data both exist and that they can be repurposed and fused to provide a training dataset.

2. Data and analytics methods

2.1. Data source and description

The GEM program internally plans sampling and re-analysis using spatial areas that are defined by the National Topographic System (NTS), which is a Natural Resources Canada standard (https://www.nrcan.gc.ca/earth-sciences/geography/topographic-information/maps/national-topographic-system-maps/9767). The southeastern Churchill Province lake-sediment dataset contains samples that are located in Labrador, encompassing parts of NTS 014D, 013M and L, 023I and J (Figs. 1 and 2). The southwestern Churchill Province and Trans-Hudson Orogen lake-sediment dataset are located in northeastern Saskatchewan, encompassing parts of NTS 064E, 074A, B, G and H (Figs. 1 and 3).

Data for a total of 6,558 samples are available across the two survey areas and was used for this proof-of-concept study, with a spatial

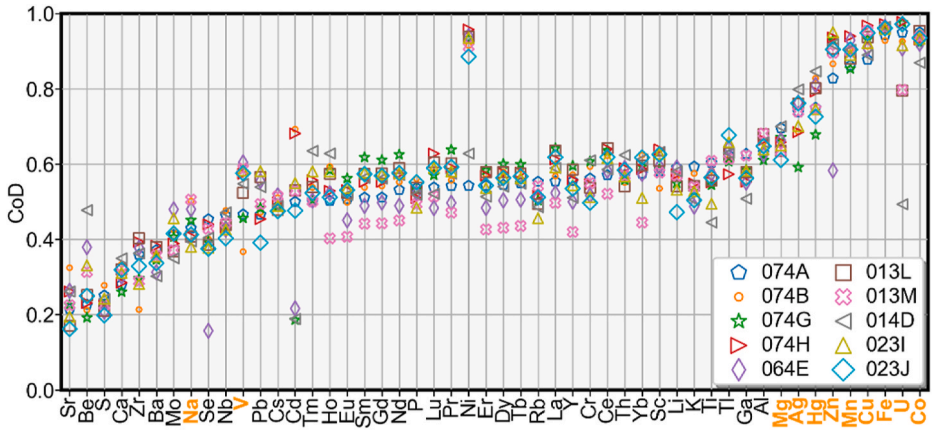

Fig. 4. Model selection performance summary – CoD (R^2) score per element for each NTS zone for the best model. Elemental order is arranged by the averaged metric score from lowest to highest. Elements that are found in all of the older datasets are shown in bold orange text (includes Pb, Mo, Ag, Hg, Mn, Ni, Zn, Co, Cu, Fe and U).

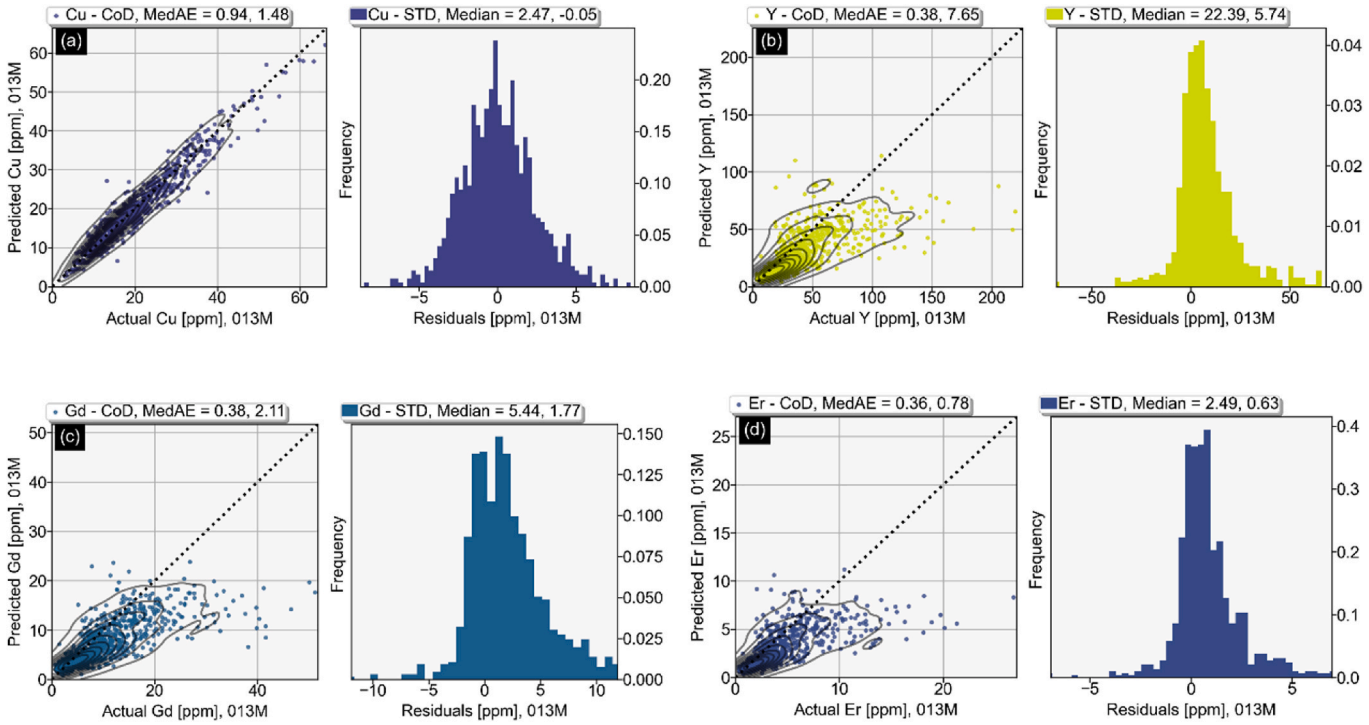

Fig. 5. Predicted versus actual for zone 013M for elements: Cu (a); Y (b); Gd (c); and Er (d). The CoD (R^2) and MedAE metric scores are shown in the scatter plots. The STD (standard deviation) and median values are shown in the histograms.

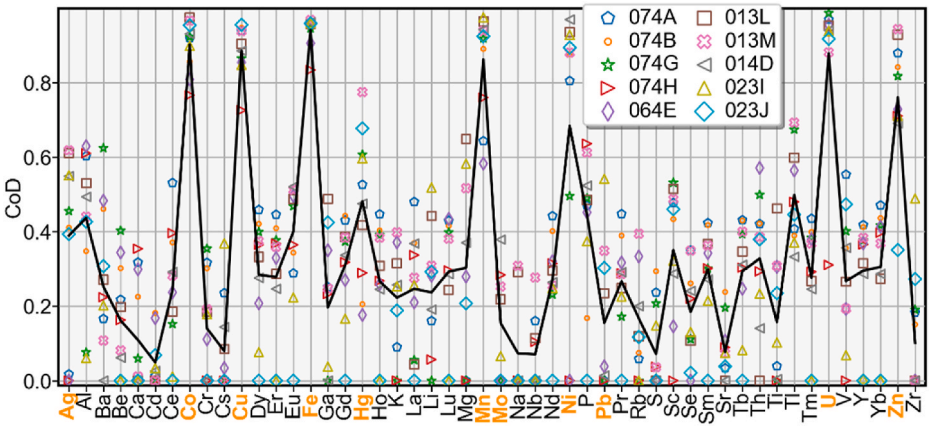

Fig. 6. Final testing performance summary – CoD (R^2) scores for all predicted elements per zone. Elements that are found in all of the older datasets are shown in bold orange text (includes Pb, Mo, Ag, Hg, Mn, Ni, Zn, Co, Cu, Fe and U).

resolution of roughly 1 sample per 13 km^2 (Friske and Hornbrook, 1991). Legacy geochemical analyses were conducted for 14–34 elements, whereas the modern analyses contain 65 elements (data descriptions are summarized in Table 1). Collected lake sediment samples for all analyses were air-dried, milled and sieved (−80 mesh). Geochemical analyses performed from 1978 to 1986 were conducted by Chemex Labs. Limited (now ALS Global) in Vancouver. All elements were determined via atomic absorption spectroscopy (AAS) with a smaller portion (up to 1,176 samples) also subjected to instrumental neutron activation analysis (INA). However, a vast majority of the analyses for all elements analyzed using INA were below the detection limit, such that the two least-affected elements – U and Fe, only contained 136 and 165 unique values. The loss on ignition (LOI) was determined via thermogravimetry, and U was determined via INA (delayed counting). Geochemical analyses performed in 1984 (OF 1643) were conducted by Barringer Magenta Ltd. In Ontario. Geochemical re-analyses performed from 2016 to 2022 used the same prepared material pulps and were conducted at Bureau Veritas in Vancouver. Pulped samples were digested using a modified aqua regia solution (1:1:1, HCl: HNO_3:H_2O). Digested samples were analyzed using an inductively coupled plasma-mass spectrometer (ICP-MS) (McCurdy et al., 2016; Bourdeau et al., 2021, 2022). Quality assurance and quality control (QA/QC) was achieved in the re-analyzed datasets by using reference materials, field duplicates and analytical duplicate samples (McCurdy and Garrett, 2016). The older datasets did not contain analyses of certified reference materials, which implies that their accuracy could not be estimated, however precision can be compared between the older and newer datasets as presented in Table 2. Digitized datasets are open source and can be obtained from the Canadian Geochemical Surveys Database at: https://geochem.nrcan.gc.ca/. The legacy and modern geochemical data was fused on a per sample basis to construct a dataset for predictive modelling.

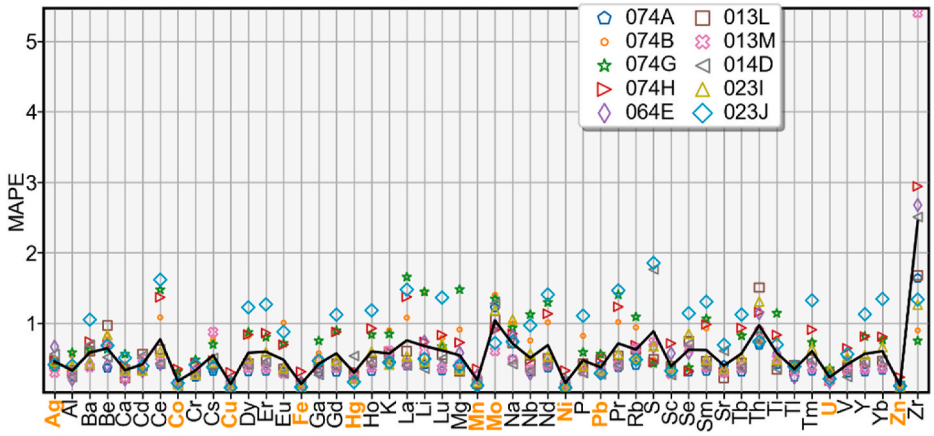

Fig. 7. Final testing performance summary – MAPE scores for all predicted elements per zone. Elements that are found in all of the older datasets are shown in bold orange text (includes Pb, Mo, Ag, Hg, Mn, Ni, Zn, Co, Cu, Fe and U).

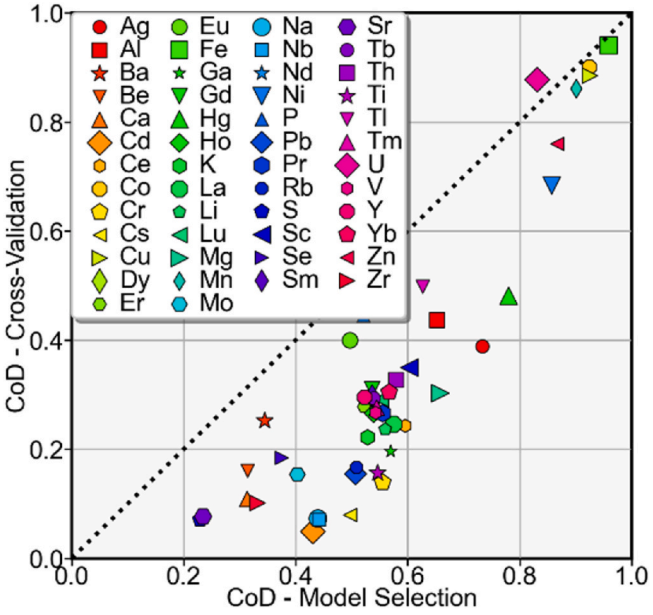

Fig. 8. Model selection versus cross-validation CoD (R^2) scores for all predicted elements. The dotted line is the 1:1 line. Points plotting below the line indicate a loss in cross-validation performance relative to model selection and vice versa.

2.2. Geological context

The data used in this study is intended to capture surficial geochemical variations of the sampled regions. Here, we provide a high-level description of the geological context to assist with the evaluation of predicted data aside from purely performance metrics. The southeastern Churchill Province is located in Labrador (Fig. 1) and has been interpreted as an elongate sliver of Archean to early-Proterozoic continental crust that was accreted between the colliding Superior (west) and North Atlantic (east) cratons during the Paleoproterozoic (James et al., 1996; James and Dunning, 2000; Wardle et al., 2002; Corrigan et al., 2018). Lithologies consist primarily of para- and ortho-gneisses, migmatites and low-to medium-metamorphic grade supracrustal and plutonic rocks. Subsequently, Mesoproterozoic-aged intrusions of variable composition (anorthosite-gabbro-troctolite, granite and syenite) intruded in the area (Hammouche et al., 2012). Notably, the Mistastin Batholith (1.4 Ga), is locally highly enriched in Rare Earth Elements [REEs], Zr, Y ± Be, Nb, U and Th (Hammouche et al., 2012). One of these zones of enrichment is hosted by the Misery Lake peralkaline syenite (1409.7 ± 1.2 Ma; David et al., 2012; Hammouche et al., 2012) and most notably by the Strange Lake peralkaline complex (1240 ± 2 Ma), which has indicated mineral resources of 278 Mt at 0.93% total REE oxide (Miller, 1990; Gowans et al., 2014; Zajac, 2015; McClenaghan et al., 2017, 2019). Glacial erosion during the Wisconsin glaciation remobilized large volumes of REE-ri ch debris towards the northeast, forming a dispersal train that can be detected in excess of 50 km down the ice (McClenaghan et al., 2017, 2019 and references therein; Zhang et al., 2022). However, the legacy data prior to the 2016 re-analysis did not contain analyses of REEs.

The southwestern Churchill Province and Trans-Hudson Orogen is located in northeastern Saskatchewan (Fig. 1). The southwestern Churchill Province is characterized by Archean gneisses, granitoids and

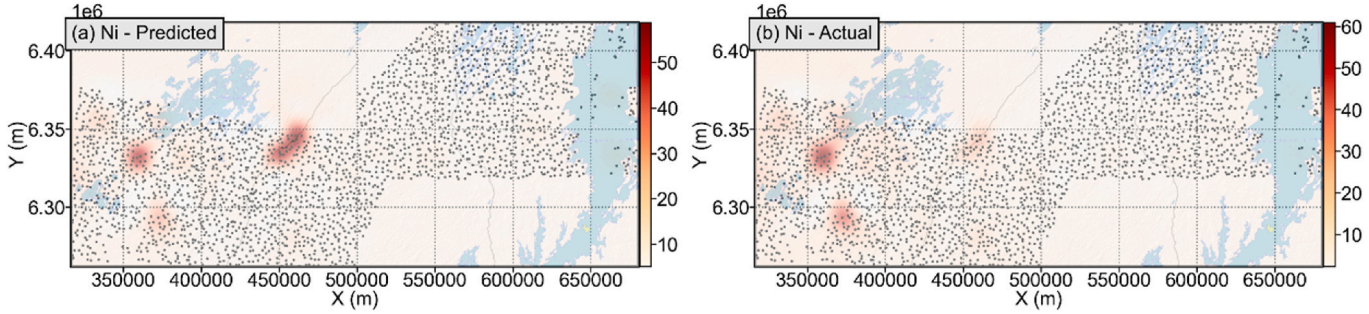

Fig. 9. Maps of the elemental concentration of Ni in the southwestern Churchill Province and Trans-Hudson Orogen region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

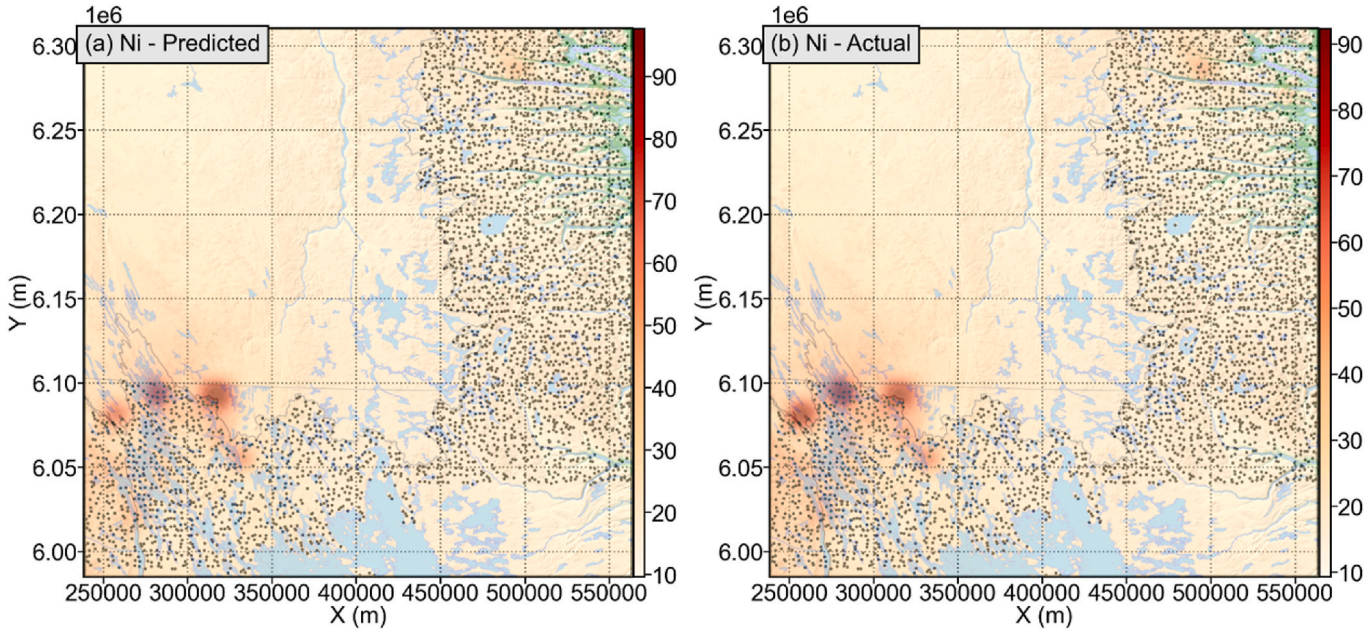

Fig. 10. Maps of the elemental concentration of Ni in the southeastern Churchill Province region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

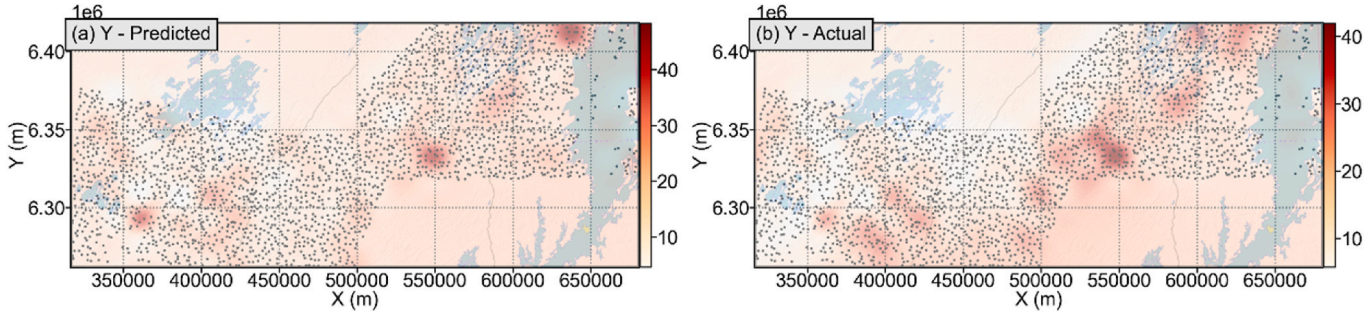

Fig. 11. Maps of the elemental concentration of Y in the southwestern Churchill Province and Trans-Hudson Orogen region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

supracrustal rocks that were overlain by Paleoproterozoic supracrustal sequences (Card et al., 2007; Eglington et al., 2013). The province hosts several mineral deposits, notably in Au (Ithingo Lake deposit) and Fe (Nyberg Lake and Ithingo Lake deposits) (Saskatchewan Geological Survey, 2018). The Churchill Province is separated from the Superior Province to the southeast by the Paleoproterozoic Trans-Hudson Orogen (1.83–1.80 Ga; Corrigan et al., 2005). The Trans-Hudson Orogen itself is a collage of juvenile supracrustal belts, continental margin sediments and re-activated rocks from the Archean Superior and Churchill provinces (Hoffman, 1988; Corrigan et al., 2005; Yeo and Delaney, 2007). A number of base-metal (e.g., Cu–Zn, Pb–Zn, Fe, Cu–Ni-Platinum Group Elements [PGE], and REEs) occurrences are found in the Trans-Hudson Orogen (Yeo and Delaney, 2007; Saskatchewan Geological Survey, 2018). These basement rocks are overlain by the Paleoproterozoic to Mesoproterozoic Athabasca Group rocks (1.75–1.5 Ga; Cumming and Krstic, 1992; Rainbird et al., 2007; Ramaekers et al., 2007). The stratigraphic areas immediately above or below the unconformity between the basement and Athabasca Group rocks is well known for Athabasca unconformity-type uranium deposits (Jeanneret et al., 2016). Legacy geochemical data for this region similarly did not contain many elements, such as the REEs.

2.3. *Machine learning workflow*

To the best of our knowledge, no studies exist that specifically examined the feasibility to predict higher dimensional geochemical data from lower dimensional and legacy geochemical data. However, there have been attempts to use machine learning and auxiliary/ancillary data to predict geochemical elemental concentrations for the purpose of mapping (e.g., Kirkwood et al., 2016) and resource estimation (Nwaila et al., 2019; Zhang et al., 2021b). For our purpose, we adopt and repurpose a ML-based workflow that is suitable for our task, which was developed by Zhang et al. (2021a, 2022). In its original formulation, it used major and minor elements to predict trace elemental concentrations, and in that context, the ML workflow was designed to reconstruct trace elemental concentrations on a per-element basis. The linear reconstruction error (prediction residual) was used to detect geochemical anomalies, primarily for a greenfield setting to delineate potential exploration targets in an unsupervised fashion. In this study, we repurpose the ML workflow to generate advanced exploration knowledge in the form of elemental concentrations of elements that have yet to be analyzed in a brownfield setting (in the sense that legacy data exists). Hence, the key predictive modelling portions of the workflow are similar to those in Zhang et al. (2021a, 2022), with the exception of the choice of features for machine learning algorithms and a removal of data

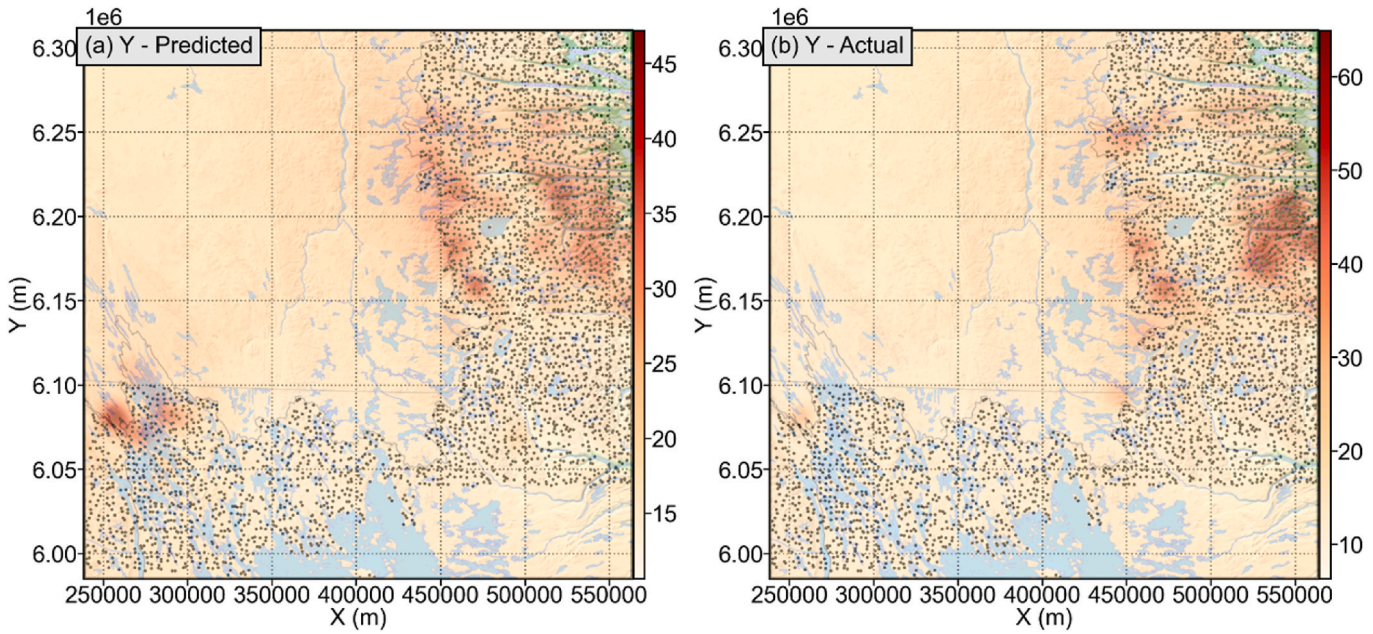

Fig. 12. Maps of the elemental concentration of Y in the southeastern Churchill Province region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

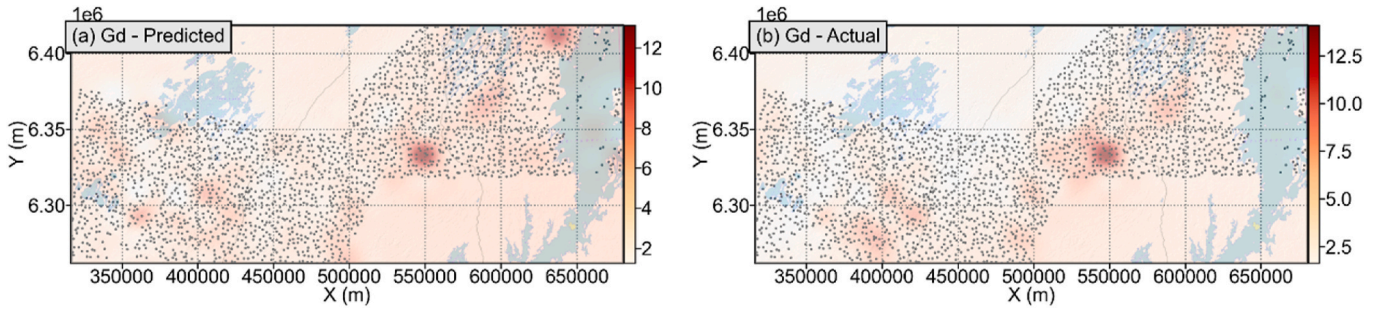

Fig. 13. Maps of the elemental concentration of Gd in the southwestern Churchill Province and Trans-Hudson Orogen region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

reconstruction-based anomaly detection. The modified ML workflow uses elements that were analyzed in the legacy analyses (analyzed between 1970 and 1980) as features and elements in the modern analyses as targets. Instead of systematically exploring all feasible machine learning algorithms, of which there are many (e.g., Zhang et al., 2021a, 2022), we choose to focus on the tuning and results of the random forest algorithm (Ho, 1995; Breiman, 1996a, b; Kotsiantis, 2014; Freund and Schapire, 1997; Sagi and Rokach, 2018). We base our selection on the fact that random forest was demonstrated to offer generally good and often the best performance for exactly this task relative to many other algorithms (Zhang et al., 2021a, 2022). For a full discussion around the random forest algorithm, and the other suitable algorithms that were explored, see Zhang et al. (2021a, 2022). For our task, some of the desirable properties of the random forest algorithm were summarized in Zhang et al. (2021a), which include a feature space that is devoid of native geometry and therefore makes no a-priori demands on the embedding space of the geochemical data. As such, it was empirically demonstrated that for the task of the prediction of trace elements, the random forest algorithm generally exhibited comparable performance across a large number of elements using both centered log ratio-transformed data and raw data, and in many cases, raw data produced better results across a range of metrics (Zhang et al., 2021a, 2022). Therefore, we choose to use raw data for predictive modelling. Model construction follows the methodology as established by Zhang et al. (2021a, 2022).

The features and targets used in this study are all the elements that are over 95% complete (<5% missing values) in the legacy and new analyses. From the legacy analyses, the features are: Hg, Fe, Mn, Ni, U, Co, Zn and Cu. The targets from the new analyses are: Cr, Dy, V, Tm, P, Th, Ni, Ti, Cs, Al, Na, Nb, Zn, Ce, Tl, Be, Co, Mg, Ag, Y, Rb, Sr, Ga, Fe, Er, Yb, U, Ca, Mn, Se, Mo, Cd, Pr, Sm, Gd, S, Hg, Nd, Eu, Ba, Sc, K, Pb, Cu, Zr, Li, Tb, Lu, La and Ho. From the datasets, all analyses are used in their raw form (either wt% or ppm). We imputed the features using a k-nearest neighbours imputer following Zhang et al. (2021a). The target elements are not imputed. Where data is missing in the targets, those data points are not used for prediction nor mapping. For model selection and tuning, we employed 4-fold cross-validation using a grid search (Table 3). Our metric of choice to perform model selection and tuning is the coefficient of determination (CoD or R^2). However, we also employed the mean absolute percentage error (MAPE; e.g., de Myttenaere et al., 2016) and the median absolute error (MedAE) to profile prediction performance. To assess predictive modelling performance in the spatial domain, we use spatial cross-validation, where the data is divided into zones along existing NTS boundaries; the target elements in each zone are predicted using data from the other zones and both variable domain performance metrics and maps are used to understand prediction performance. Segmentation by NTS zones is similar to how the GEM program plans surveys and re-analysis campaigns. Therefore,

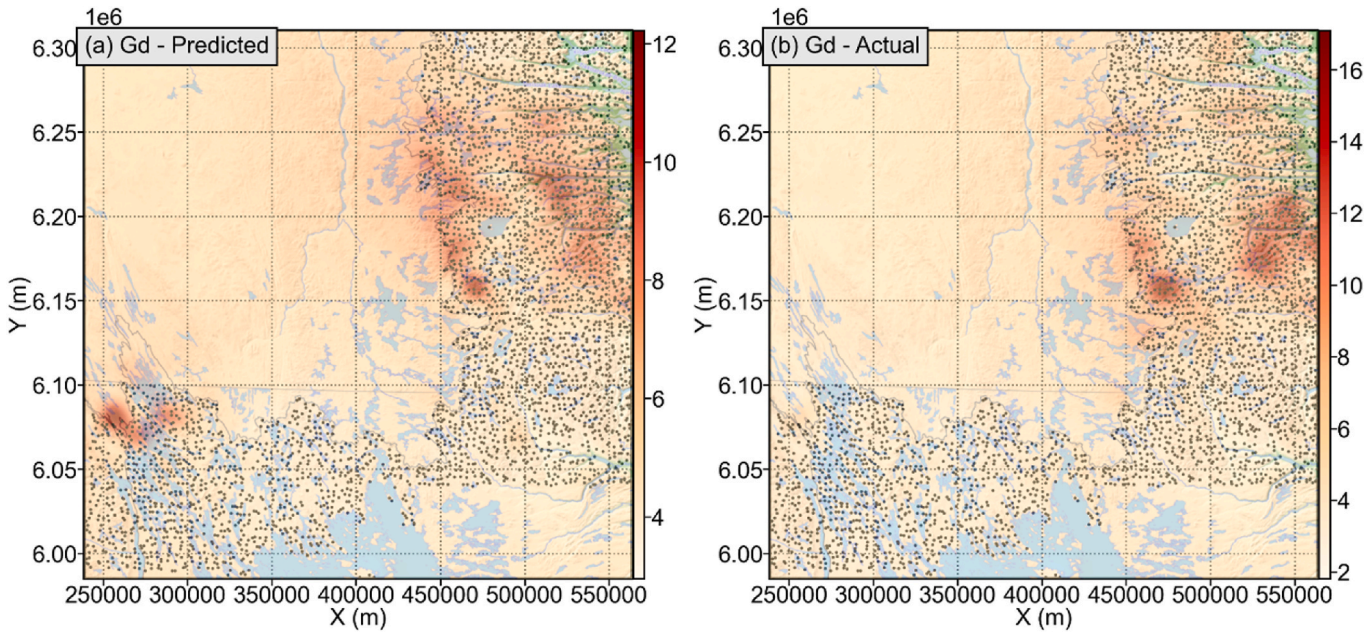

Fig. 14. Maps of the elemental concentration of Gd in the southeastern Churchill Province region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

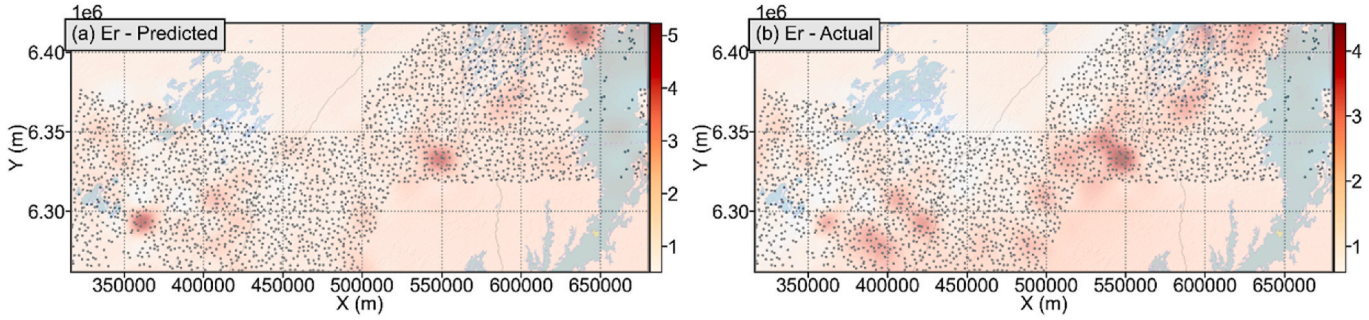

Fig. 15. Maps of the elemental concentration of Er in the southwestern Churchill Province and Trans-Hudson Orogen region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

spatial cross-validation using this method resembles how predictive modelling ahead of re-analyses would be operationalized.

3. Results

Model selection results demonstrate that the best-predicted elements are the elements that overlap between the legacy and modern geochemical analyses. In contrast, the other elements are predicted with a variable CoD (R^2) depending on the element and the NTS zones (Fig. 4). Some elements are predicted poorly (CoD (R^2) < 0.3; e.g., Sr and S). Some zones consistently predict worse across some elements, such as zone 013M and elements: Ho; Eu; Sm; Gd; Nd; Pr; Er; Dy; Tb; Y; Ce and Yb. This zone is known to contain some rare earth element-bearing geochemical anomalies associated with the Mistastin Batholith (McClenaghan et al., 2017, 2019 and references therein). As such, a reduction in prediction performance is expected, and indeed, data reconstruction performance was used to identify mineralization anomalies (Zhang et al., 2021a, 2022). Results from testing using spatial cross-validation show that prediction of various elements at the NTS zone level is generally possible. In particular, predicted concentrations of the rare earth elements in zone 013M tend toward under-prediction (Fig. 5). This same behaviour was observed in Zhang et al. (2022).

The prediction performance observed through spatial cross-validation is roughly congruent with those of model selection (Fig. 6). Without using spatial cross-validation (e.g., data splitting is random from all NTS zones), the performance is generally substantially closer to those of model selection (Fig. 6). In general, there is a reduction in overall performance and a tendency of some zones to predict better than others across a range of elements, such as zone 074A for most of the predicted rare earth elements (Fig. 6). Zone 023J was the least predictable on average across many elements, such as the rare earth elements (Fig. 6). This is because zone 023J contains an abundance of data that is chemically quite dissimilar to those in the other zones. Therefore, models trained on data from zones exclusive of zone 023J did not generalize well to zone 023J. This behaviour is also observed using the MAPE metric (Fig. 7). For most of the elements except for Th, Mo and Zr, the weighted averages of MAPE for all other elements are below 1 (Fig. 7). This indicates that the median error for most elements is, on average lower than their concentration values. To better understand the performance expectations, such as model generalizability, we examined the change in the CoD (R^2) scores between model selection and cross-validation phases (Fig. 8). Some elements (e.g., Cd, Cs, Cr, Nb and Na) exhibited a relatively significant loss in performance in cross-validation relative to model selection (Fig. 8).

Most maps produced using predicted elemental concentrations exhibit variably resemblant patterns to the actual elemental

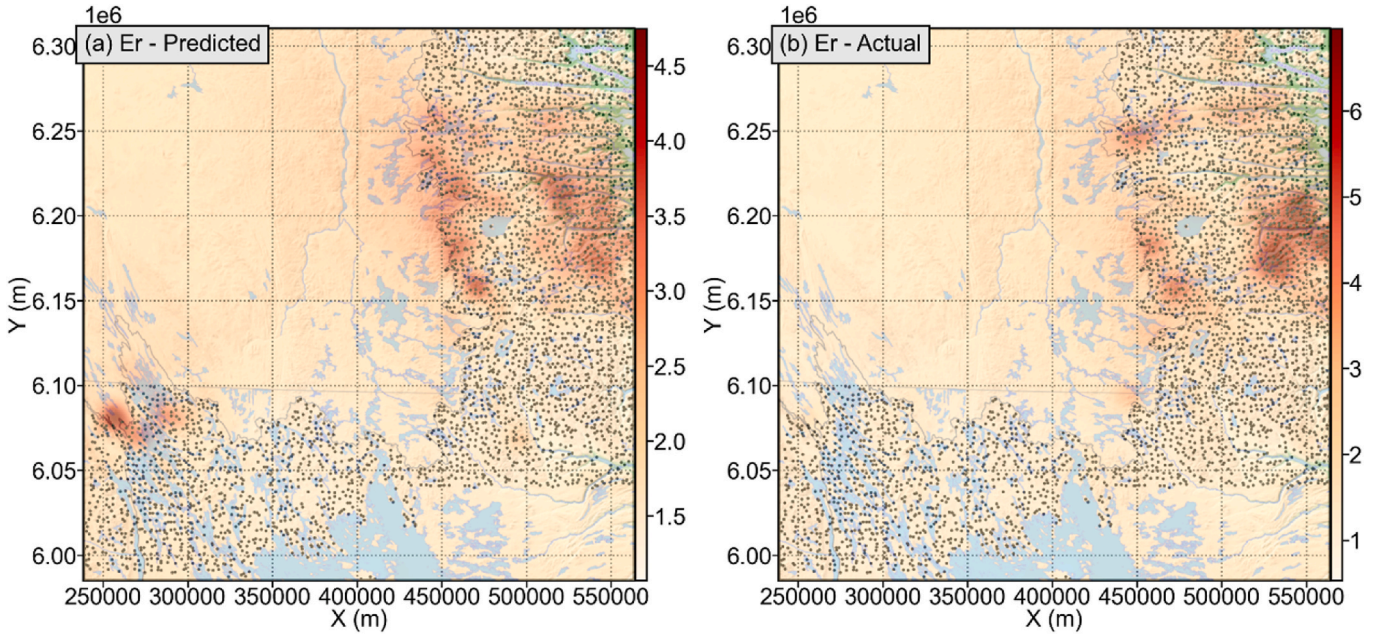

Fig. 16. Maps of the elemental concentration of Er in the southeastern Churchill Province region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

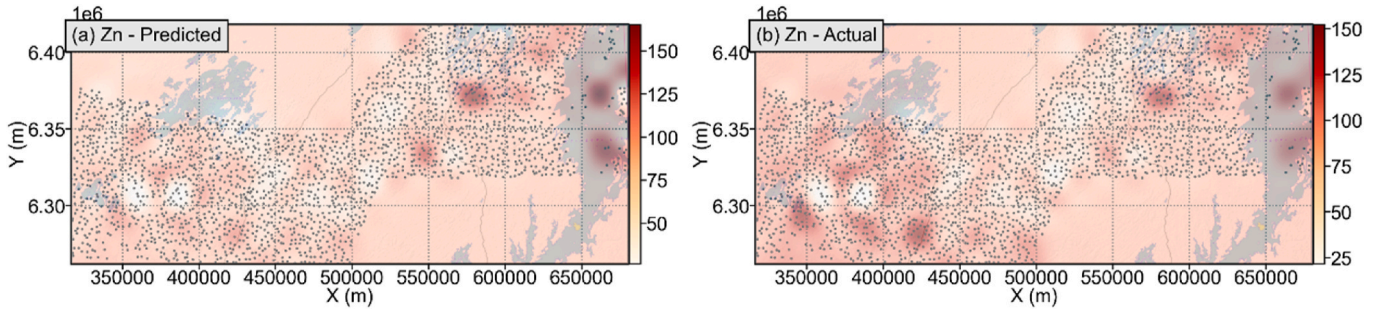

Fig. 17. Maps of the elemental concentration of Zn in the southwestern Churchill Province and Trans-Hudson Orogen region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

concentration maps (Figs. 9–20). The degree of spatial concordance is high for elements that were contained both in the legacy and modern datasets, such as Ni (Figs. 9 and 10). In fact, in the Strange Lake region (southeastern Churchill Province), the maps of the predicted and actual elemental concentrations are qualitatively indistinguishable (Figs. 9 and 10). In the southwestern Churchill Province and Trans-Hudson Orogen region, the Ni maps are qualitatively similar, but the prediction suggests a higher concentration in the central region of the map (Fig. 9). However, predicting concentrations of elements that already exist in legacy data is less interesting than predicting currently unknown elemental concentrations. For rare earth elements, the degree of spatial concordance is generally good qualitatively because zones known to be enriched in some elements are predicted to be as such (Figs. 11–16). In particular, in the southeastern Churchill Province region, locations of known rare earth element-bearing anomalies such as the Strange Lake peralkaline complex and Misery Lake peralkaline syenite (Mistatin Batholith), as well as their dispersal trains are predicted to be enriched in those elements (Figs. 12, 14 and 16). For base metals, the degree of spatial concordance depends on the elemental overlap between the legacy and modern datasets and, in some cases, such as for Zn and Pb, the spatial concordance is quite good in the southeastern Churchill Province region (Figs. 18 and 20), but poorer in portions of the southwestern Churchill Province and Trans-Hudson Orogen region (mainly the west half; Figs. 17 and 18).

Since the purpose of the legacy and re-analyzed geochemical data is primarily for regional mapping, the assessment of prediction performance can also occur in the image domain. The image domain contains the gridded data and the blocks in the grid have been estimated using interpolation. In contrast to performance assessments of raw predicted data in the variable domain, the gridding and interpolation process, such as that employed in this proof-of-concept, is more suitable to understand the performance of our approach, where mapping is an anticipatable downstream activity (e.g., prospectivity mapping). However, the gridding and interpolation process introduces two additional mechanisms that may alter performance relative that of the raw predicted data. One mechanism is the error incurred by interpolation, which strongly depends on the interpolation approach used. The other mechanism is the spatial averaging effect of estimating a single block from multiple adjacent supporting points. This is expected to increase the performance of the predicted data when measured in the image domain, if the predicted data exhibits a residual that is substantially symmetrical around zero (such that averaging over a random sample tend to converge to the mean, or nearer to zero) and some amount of spatial correlation exists (otherwise mapping is unwarranted). The symmetry of the residuals is generally observable but is more significant for some elements than others (e.g., Cu in Fig. 5 relative to the other elements). In general, there is a significant improvement of all metrics when measured in the image domain relative to that of the raw predicted data (Figs. 21–23). The

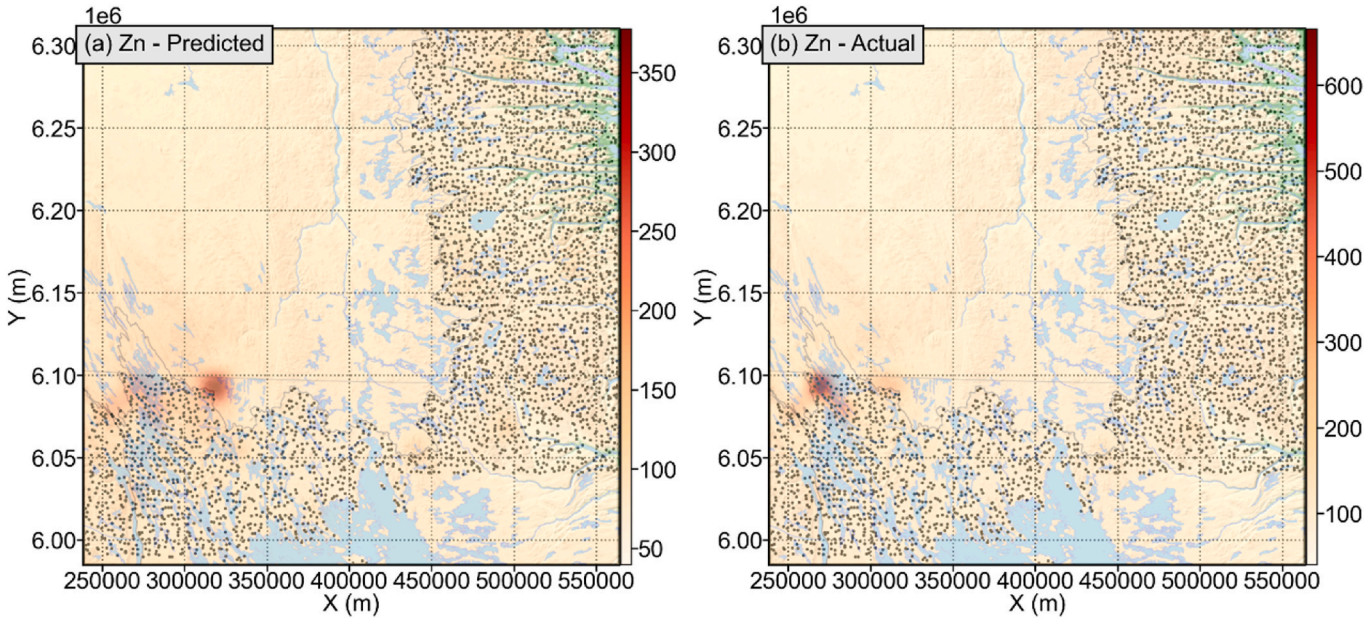

Fig. 18. Maps of the elemental concentration of Zn in the southeastern Churchill Province region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

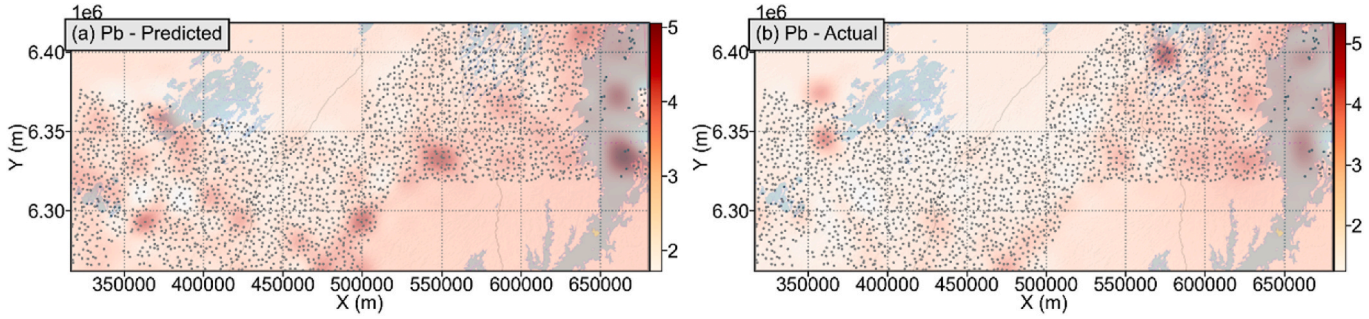

Fig. 19. Maps of the elemental concentration of Pb in the southwestern Churchill Province and Trans-Hudson Orogen region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

difference can be measured through the relative difference (image domain score minus raw predicted-data score, divided by the raw predicted-data score). The relative difference for: the CoD (R^2) metric is about 42.21%; the MAPE metric is about −69.24%; and the MedAE metric is about −37.83%. This implies that the MAPE metric experienced the most improvement, followed by the CoD (R^2) metric. Furthermore, for the MAPE and MedAE metrics and with the exception of Mn (Figs. 22 and 23), all elements systematically improved in the image domain relative to the same metrics evaluated on the raw predicted data. This implies that predicted geochemical data is generally suitable for mapping.

4. Discussion

This study shows that it is possible to predict elemental concentrations for a range of elements in areas where legacy geochemical data exist, but the corresponding modern data is absent. Cross-validation can be used to anticipate prediction performance using a variety of metrics in the variable domain (Figs. 6–8). Elements to be predicted that are present in training data generally predict better in any given area than elements that are non-existent in training data (Figs. 6 and 7). This is no surprise, as information regarding such an element is more direct than the inference of presently unanalyzed elements. Therefore, predicted elemental concentration maps vary quantitatively in their accuracy (Figs. 9–20). High CoD (R^2) scores in the variable domain (predicted versus actual concentrations for the raw predicted data) typically translated into predicted concentration maps that were consistent to nearly identical with the actual concentration maps (Figs. 9–20, compared with metric scores in Figs. 21–23). In the image domain (gridded and interpolated predicted data versus that of the ungridded data), all metric scores were generally substantially improved relative to that in the variable domain (Figs. 21–23). This highlights the positive impact of spatial correlation and local averaging, which seems to generally favour the rendering of maps using predicted data in our case. As the CoD (R^2) score (in all domains) decrease, there is a general loss of spatial concordance between the predicted and actual elemental maps (e.g., Ni in Fig. 10 compared with Pb in Fig. 19). At lower values of CoD (R^2), two key mechanisms contribute to discrepancies between the predicted and actual concentration maps. These are: (1) under- or over-prediction of concentration hotspots (e.g., Fig. 18, the western half of the map) and; (2) prediction of concentration hotspots that are not present in the actual concentration maps (e.g., Fig. 16, the lower western portion of the map). However, (2) appears to be less common in our results and in cases where seemingly false-positive hotspots are present for some elements, the locations are important for other similar elements (e.g., Fig. 16, the lower western portion of the map compared with Fig. 12b, the same region showing a small enrichment in Y). In this sense, much of the occurrences of (2) are better characterized as

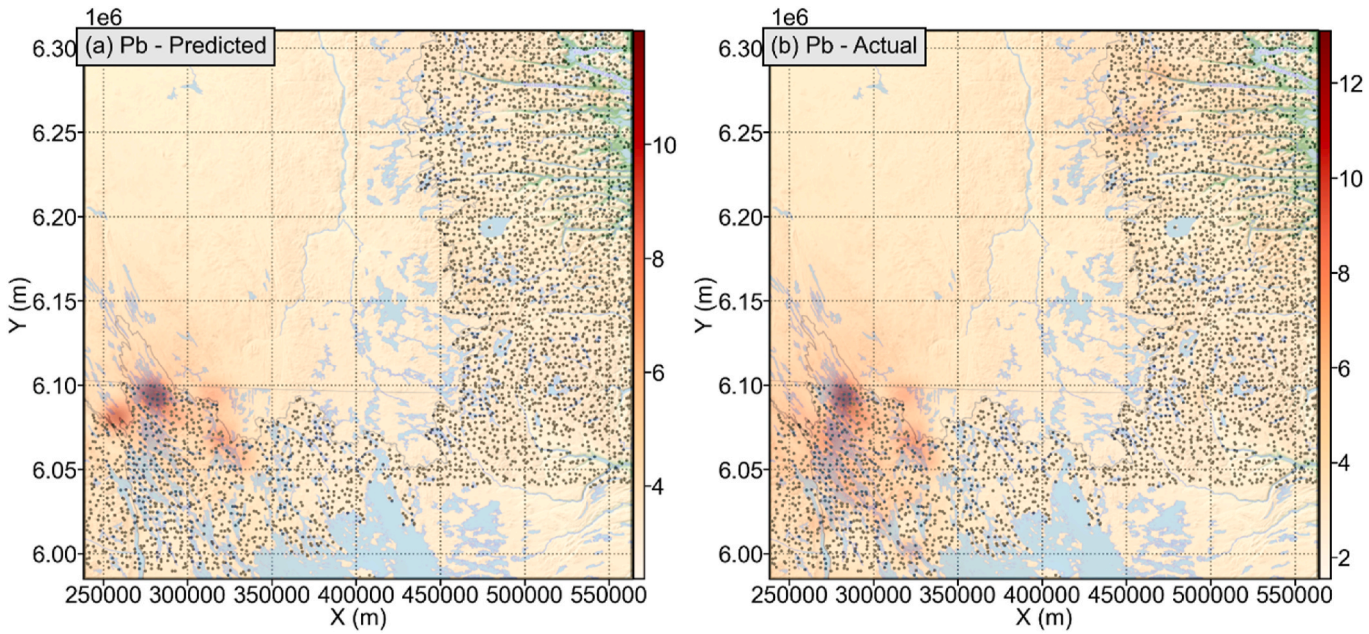

Fig. 20. Maps of the elemental concentration of Pb in the southeastern Churchill Province region were created from (a) predicted concentrations (ppm), and (b) actual concentrations (ppm).

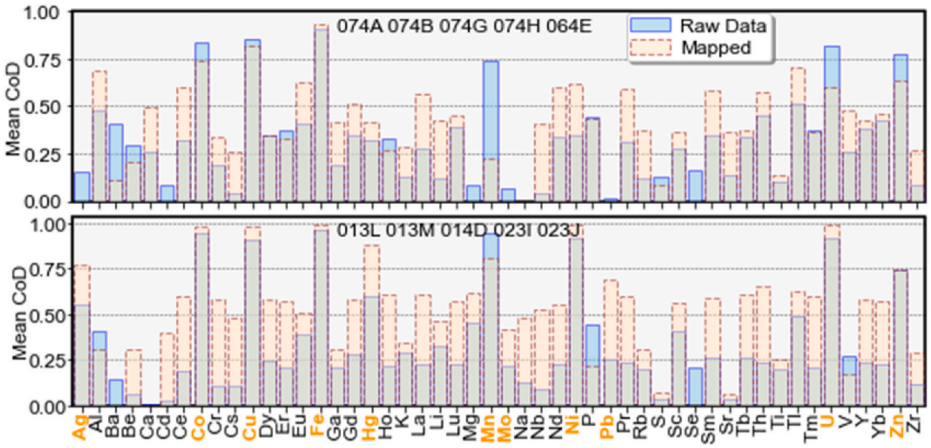

Fig. 21. Comparison of the CoD (R^2) metric scores on a per-element basis between the raw predicted data and the mapped data. Elements that are found in all of the older datasets are shown in bold orange text (includes Pb, Mo, Ag, Hg, Mn, Ni, Zn, Co, Cu, Fe and U).

multivariate/multi-elemental concentration hotspots, for which some of the elements are very low in concentration. In any case, it is clear that because the predicted concentration maps generally bear a spatial resemblance to the actual concentration maps, such maps could be used to plan re-analyses or even additional sampling campaigns (Figs. 9–20). In addition, the predicted and actual hotspots are nearly always of the same order of magnitude, which is quantitatively supported by the MAPE metric scores (nearly always less than 100%, see Fig. 7), which implies that advanced targeting decisions could be made using predicted concentration maps, in lieu of waiting for the areas to be re-analyzed. This may be particularly important for time-sensitive elements, such as those belonging to the critical raw materials. In any case, multivariate concentration maps are likely to be generally more robust than single elemental maps, as stacking of several maps would attenuate spatial disagreements amongst the elements and enhance hotspots that are common across the elements. For much of the rare earth elements, their geochemical behaviour is comparable. Therefore, grouping them together to generate final multivariate elemental maps is a good use of the prediction results. This will be useful for advanced exploration of areas that do not feature modern multi-element geochemical analyses, particularly where an entire group of elements (e.g., the rare earth elements) are missing but are important operationally.

Prediction performance is highly dependent on the quality of the training data and particularly its relationship with the target region. In general, to maximize prediction performance, the training data should encompass data that is as representative of the target region as possible to increase model generalizability (e.g., the class imbalance problem). Additionally, for algorithms that cannot extrapolate beyond the target's numerical range (such as the random forest algorithm and in general, tree-based algorithms), the training data should ideally contain ranges of data along each feature and target that entirely overlap with those of the deployment dataset. This may be less relevant for qualitative uses of the prediction outputs (e.g., a geochemical hotspot is found, but its exact concentration is not important). In the context of geochemical data, this implies that the compositions of the testing samples should be covered in abundance in the training samples. In the case of lake sediment samples, this further implies that the parent rocks from which the sediment was derived should be adequately covered in the training dataset. In other

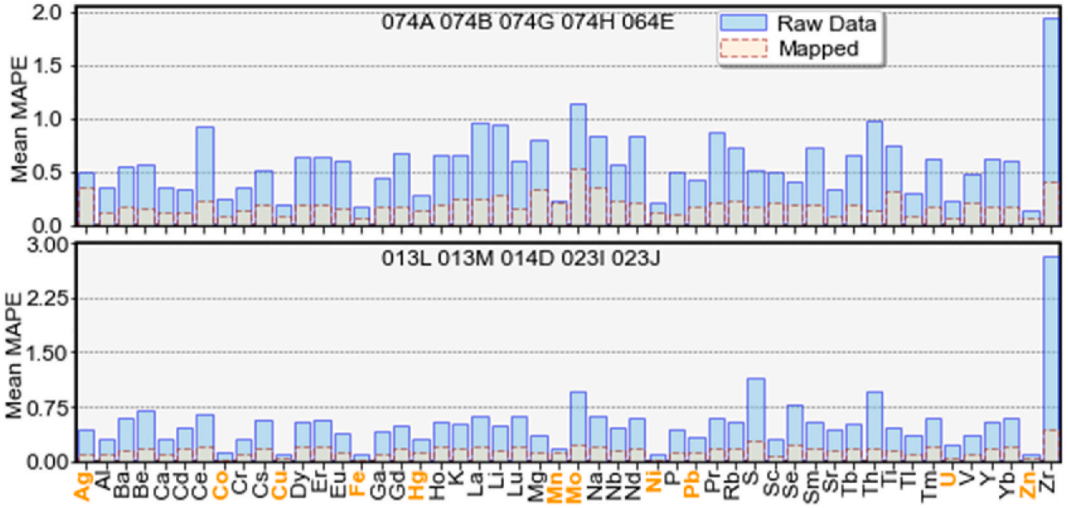

Fig. 22. Comparison of the MAPE metric scores on a per-element basis between the raw predicted data and the mapped data. Elements that are found in all of the older datasets are shown in bold orange text (includes Pb, Mo, Ag, Hg, Mn, Ni, Zn, Co, Cu, Fe and U).

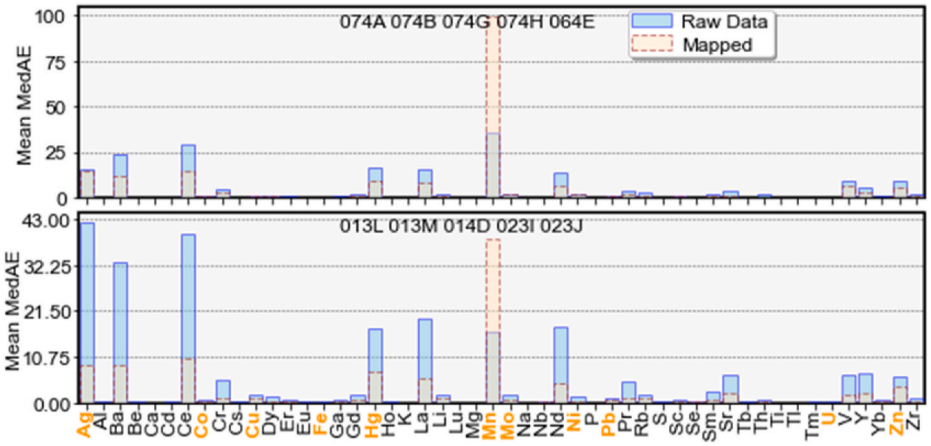

Fig. 23. Comparison of the MedAE metric scores on a per-element basis between the raw predicted data and the mapped data. Elements that are found in all of the older datasets are shown in bold orange text (includes Pb, Mo, Ag, Hg, Mn, Ni, Zn, Co, Cu, Fe and U).

words, prediction and training using contrasting compositions of samples, e.g., training on sediment samples derived from parental igneous rocks and testing on sediment samples derived from conglomerates, is unlikely to produce optimal performance. There are several classes of factors that can contribute to significant losses in prediction performance in this manner: (1) scientific factors; and (2) data factors. Scientific factors include the abundance of relationships (e.g., types of minerals in samples) between the target elements and the features. For example, an element is found in spatially disparate stoichiometries in samples that vary at a spatial scale roughly the size of the NTS zones (a type of mineral spatial non-stationarity through e.g., the distribution of parent lithologies). In this case, in the worst case, there are unique relationships between the targets and features within each zone and generalization to unknown zones is critically dependent on having sufficient types of relationships captured in the training data. Data factors include under-sampling of key relationships for some elements and class imbalance issues. To some extent, all of these issues can be reduced in their impacts through increasing the diversity of the training data through the availability of more re-analyses. Sourcing more data through other sources external to GEM, such as provincial, academic and private industry data may also be a means to augment the diversity of the training data. However, discrepancies between sample collection, preparation and analysis methods should ideally be taken into account and depending on the impact of the discrepancies, data corrections may have to be carried out prior to integrating such data into the training data.

For the optimization of prediction performance, geological and geochemical knowledge of target areas may be vital in creating effective training datasets. However, as the amount of data increases due to more data becoming modernized through re-analyses, then the geological and geochemical knowledge should become less relevant, and a more or complete data-driven approach could be used instead. We anticipate that as more training data becomes available through re-analyses, model generalizability and prediction performance should generally improve. The high degree of comparability between datasets through standardized practices in data generation (e.g., sampling, instrumentation and analytical standards) makes this type of approach possible. Although Zhang et al. (2021a) demonstrated that prediction of elemental concentrations are generally reliable even using a secondary dataset with unknown and impossible-to-level data, further investigation of the impacts of systematic data levelling issues may be necessary (although beyond the scope and data availability of this study). Data levelling in our application is less important or completely irrelevant between the features and targets (e.g., levelling between strictly legacy and modern datasets) but is more important for building a robust training database (e.g., levelling between legacy datasets). Thankfully, the open data

nature of the analyses facilitates the comprehension of sampling and analytical methodologies. This should be as applicable for GEM data as it is for similar data, wherever conditions are met (e.g., fusable legacy and modern datasets). The availability of common standards enables data levelling in the future, where its effects are intolerable. Where the predicted elemental concentration maps are used qualitatively (e.g., to plan re-analysis or sampling campaigns) instead of quantitatively (e.g., to substitute for re-analyses in an exploration campaign or to be used with scientifically reductive methods downstream), quantitative matching of data levels may be unnecessary depending on the application context and the extent of its impacts.

The ability to generate a-priori knowledge regarding previously unknown elemental concentrations in new areas is the most significant finding of this study. We demonstrated that where previous analyses exist in the form of legacy data, it is possible to anticipate the elemental concentrations of many unanalyzed elements with sufficient accuracy in the spatial domain to be widely applicable. Where exploration programs were sufficiently regimental to have had standardized sampling and analysis methodology, and some implementation of a data repository that facilitates the use of particularly legacy data, the legacy data and modern data are a key program asset that enables predictive operationalization. Our finding implies that re-analysis campaigns and exploration for time-sensitive elements (e.g., critical raw materials) could become prioritized based on anticipated elemental concentrations and current exploration priorities. For example, in the search for rare earth elements, it is possible to sequentially target areas (both for further downstream prospectivity mapping or for detailed field study and sampling) based on a rank-order of their anticipated rare earth elemental concentrations. This effectively reduces the time to discover new mineral deposits by reshaping the exploration pipeline to become more predictive in a target- and data-driven manner. In this case, geological knowledge of target areas is an asset but not a requirement, as the decision to target is data-driven. However, geoscientific knowledge is indispensable for interpreting and verifying relevant geochemical hotspots in predicted elemental maps. In some situations, aside from exploration prioritization, advanced geochemical data through predictive modelling may be the only form of modern geochemical data, and such data is better than legacy data alone. This would be the case in circumstances satisfying one or more of: (1) the time to modernize data in an area of interest is too long for a practical purpose; (2) there is insufficient material remaining for another re-analysis and/or re-sampling is no longer possible (e.g., due to changes in land use; geopolitics; environmental, social and governance factors); (3) there is no additional resource to execute another re-analysis campaign; (4) probable rather than known concentrations are adequate for downstream uses of data; and (5) re-analysis is contingent on more geochemical information becoming available in target areas.

5. Conclusion

Predictive data generation can play a pivotal role in the modernization of exploration programs by providing probable data in advance of actual data. We have demonstrated that by using legacy data, modern data and leveraging the extensive standardization of data and its generation processes within existing geochemical survey programs, it is possible to generate probable and often quantitatively or qualitatively reliable modern data. In this manner, our proposed method not only extracts additional value from otherwise outdated (legacy) data, but also serves to assist with both upstream data generation and downstream data usage. For exploration geochemistry, the rate-limiting step is data generation. This context is mismatched with the data-hungry and automation capacity of data-driven techniques, which are becoming key downstream consumers of exploration data and can quickly exhaust existing data. In this case, our method serves to enable data users, such as prospectivity mappers, with valuable and timely information that otherwise may be completely absent or may be untimely. This would enable the generation of advanced probable knowledge of target areas. An effect of this advanced knowledge is that it can potentially make future upstream exploration activities, such as data generation more targeted and fit-for-purpose for downstream data analysis. This would obviously facilitate timely resource discoveries, which is important for critical raw materials. Additionally, focusing on targeted areas may lead to the detection of fainter geochemical anomalies, which increases the probability of resource discovery. Although knowledge derived using predicted data is unlikely to be as accurate or selective as those using actual data, the accuracy of predicted data is only likely to improve as more modern data becomes available for training. As data is becoming the currency of digital societies, then in the manner of financial discounting, advanced knowledge available now may be worth more than more accurate knowledge that is available later.

Funding

This study was supported by a Department of Science and Innovation (DSI)-National Research Foundation (NRF) Thuthuka Grant (Grant UID: 121973), and DSI-NRF CIMERA.

Declaration of competing 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.

Acknowledgements

The data used in this study are open source and are hosted at the Canadian Database of Geochemical Surveys Portal: https://geochem.nrcan.gc.ca/cdogs/content/main/home_en.htm, an open-source portal managed by Natural Resources Canada (NRCan). The authors would also like to thank two 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.

References

Balaram, V., 2021. Current and emerging analytical techniques for geochemical and geochronological studies. Geol. J. 56, 2300–2359. https://doi.org/10.1002/gj.4005.

Bourdeau, J.E., Day, S.J.A., Zhang, S.E., 2022. Regional Lake Sediment Geochemical Data from Northeastern Saskatchewan (NTS 064-E, 074-A and H): Re-analysis and QA/QC Evaluation, vol. 8871. Geological Survey of Canada, Open File, p. 19. https://doi.org/10.4095/329875.

Bourdeau, J.E., McCurdy, M.W., Day, S.J.A., Garrett, R.G., 2021. Regional Lake Sediment Geochemical Data from North-Central Saskatchewan (NTS 074-A, B, G, and H): Re-analysis Data and QA/QC Evaluation, vol. 8837. Geological Survey of Canada, Open File, p. 14. https://doi.org/10.4095/329204.

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.

Card, C.D., Pană, D., Portella, P., Thomas, D.J., Annesley, I.R., 2007. Basement rocks to the Athabasca basin, saskatchewan and alberta. In: Jefferson, C., Delaney, G. (Eds.), EXTECH IV: Geology and Uranium EXploration TECHnology of the Proterozoic Athabasca Basin, Saskatchewan and Alberta, vol. 588. Geological Survey of Canada, Bulletin, pp. 69–87. https://doi.org/10.4095/223745.

Cohen, D.R., Kelley, D.L., Anand, R., Coker, W.B., 2010. Major advances in exploration geochemistry, 1998-2007 Geochem. Explor. Environ. Anal. 10, 3–16. https://doi.org/10.1144/1467-7873/09-215.

Corrigan, D., Hajnal, Z., Németh, B., Lucas, S.B., 2005. Tectonic framework of a Paleoproterozoic arc–continent to continent–continent collisional zone, Trans-Hudson Orogen, from geological and seismic reflection studies. Can. J. Earth Sci. 42, 421–434. https://doi.org/10.1139/e05-025.

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.

Council for Geosciences, 2022. Annual Report, 2021/22. Council for Geosciences, Replublic of South Africa, p. 162.

Cumming, G.L., Krstic, D., 1992. The age of unconformity-related uranium mineralization in the Athabasca Basin, northern Saskatchewan. Can. J. Earth Sci. 29, 1623–1639. https://doi.org/10.1139/e92-128.

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, p. 33. RP 2012-01.

de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F., 2016. Mean absolute percentage error for regression models. Neurocomputing 192, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114.

Eglington, B.M., Pehrsson, S.J., Ansdell, K.M., Lescuyer, J.L., Quirt, D., Milesi, J.P., Brown, P., 2013. A domain-based digital summary of the evolution of the Palaeoproterozoic of North America and Greenland and associated unconformity-related uranium mineralization. Precambrian Res. 232, 4–26. https://doi.org/10.1016/j.precamres.2013.01.021.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Friske, P.W.B., Hornbrook, E.H.W., 1991. Canada's national geochemical reconnaissance programme. In: Transactions of the Institution of Mining and Metallurgy, Section B. Applied Earth Science, vol. 100, pp. B47–B56.

GEM-GeoNorth, 2022. Geoscientific Research in Canada's North: GEM-GeoNorth from. https://www.nrcan.gc.ca/earth-sciences/resources/federal-programs/geomapping-energy-minerals/18215. (Accessed 9 May 2022).

Geological Survey of Canada, 2022. Geological survey of Canada: report on results and delivery 2020-2021. Natural Resources Canada 137. https://doi.org/10.4095/328919.

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

Hammouche, H., Legouix, C., Goutier, J., Dion, C., 2012. Géologie de la région du lac Zeni. Ministère des Ressources Naturelles Québec, p. 35. RG 2012-02.

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.

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.

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.

Jeanneret, P., Gondlaves, P., Durand, C., Trap, P., Marquer, D., Quirt, D., Ledru, P., 2016. Tectono-metamorphic evolution of the pre-athabasca basement within the wollaston-mudjatik transition zone, saskatchewan. Can. J. Earth Sci. 53, 231–259. https://doi.org/10.1139/cjes-2015-0136.

Kotsiantis, S.B., 2014. Integrating global and local application of naive bayes classifier. Int. Arab J. Inf. Technol. 11 (3), 300–307.

Kirkwood, C., Cave, M., Beamish, D., Grebby, S., Ferreira, A., 2016. A machine learning approach to geochemical mapping. J. Geochem. Explor. 167, 49–61. https://doi.org/10.1016/j.gexplo.2016.05.003.

Lebel, D., 2020. Geological Survey of Canada 8.0: mapping the journey towards predictive geoscience. In: Hill, P.R., Lebel, D., Hitzman, M., Smelror, M., Thorleifson, H. (Eds.), The Changing Role of Geological Surveys. The Geological Society, London, Special Publication, pp. 49–69. https://doi.org/10.1144/SP499-2019-79.

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.

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, vol. 8240. Geological Survey of Canada, Open File, p. 50. https://doi.org/10.4095/306239.

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, Open File 8026, p. 14. https://doi.org/10.4095/298834.

McCurdy, M.W., Garrett, R.G., 2016. Geochemical data quality control for soil, till and lake and stream sediment samples. Geological Survey of Canada, Open File 7944, 40. https://doi.org/10.4095/297562.

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.

Nwaila, G.T., Zhang, S.E., Frimmel, H.E., Manzi, M.S.D., Dohm, C., Durrheim, R.J., Burnett, M.S., Tolmay, L.C., 2019. Local and target exploration of conglomerate-hosted gold deposits using machine learning algorithms: a case study of the Witwatersrand gold ores, South Africa. Nat. Resour. Res. 29, 135–159. https://doi.org/10.1007/s11053-019-09498-1.

Rainbird, R.H., Stern, R.A., Rayner, N., Jefferson, C.W., 2007. Age, provenance, and regional correlation of the Athabasca Group, Saskatchewan and Alberta, constrained by igneous and detrital zircon geochronology. In: Jefferson, C., Delaney, G. (Eds.), EXTECH IV: Geology and Uranium EXploration TECHnology of the Proterozoic Athabasca Basin, Saskatchewan and Alberta, vol. 588. Geological Survey of Canada, Bulletin, pp. 193–209. https://doi.org/10.4095/223761.

Ramaekers, P., Jefferson, C.W., Yeo, G., Collier, B., Long, D.G.F., Catueanu, O., Bernier, S., Kupsch, B., Post, R., 2007. Revised geological map and stratigraphy of the Athabasca group, saskatchewan and alberta. In: Jefferson, C., Delaney, G. (Eds.), EXTECH IV: Geology and Uranium EXploration TECHnology of the Proterozoic Athabasca Basin, Saskatchewan and Alberta, vol. 588. Geological Survey of Canada, Bulletin, pp. 155–191. https://doi.org/10.4095/223754.

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.

Saskatchewan Geological Survey, 2018. Resource Map of Saskatchewan, 2018 Edition. Saskatchewan Ministry of the Economy, Saskatchewan Geological Survey, Miscellaneous. Report 2018-1.

Van Rooyen, R.C., Elsenbroek, J.H., Balt, T.O., 2004. Regional Geochemical Map Series, 1:1,000,000, vol. 2528. Council for Geosciences, Pretoria, p. 2.

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.

Yeo, G.M., Delaney, G., 2007. The wollaston supergroup, stratigraphy and metallogeny of a paleoproterozoic wilson cycle in the trans-hudson orogen, saskatchewan. In: Jefferson, C., Delaney, G. (Eds.), EXTECH IV: Geology and Uranium EXploration TECHnology of the Proterozoic Athabasca Basin, Saskatchewan and Alberta, vol. 588. Geological Survey of Canada, Bulletin, pp. 89–117. https://doi.org/10.4095/223746.

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., Bourdeau, J.E., Nwaila, G.T., Corrigan, D., 2022. Towards a fully data-driven prospectivity mapping methodology: a case study of the Southeastern Churchill Province, Québec and Labrador. Artif. Intell. Geosci. 2, 128–147. https://doi.org/10.1016/j.aiig.2022.02.002.

Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Ashwal, L.D., 2021a. 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. 2, 60–75. https://doi.org/10.1016/j.aiig.2021.11.002.

Zhang, S.E., Nwaila, G.T., Tolmay, L.C., Frimmel, H.E., Bourdeau, J.E., 2021b. Integration of machine learning algorithms with gompertz curves and kriging to estimate resources in gold deposits. Nat. Resour. Res. 30, 39–56. https://doi.org/10.1007/s11053-020-09750-z.