Titre: Title:	Quantifying uncertainty and improving prospectivity mapping in mineral belts using transfer learning and Random Forest: A case study of copper mineralization in the Superior Craton Province, Quebec, Canada
Auteurs: Authors:	Dany Lauzon, & Erwan Gloaguen
Date:	2024
Type:	Article de revue / Article
Référence: Citation:	Lauzon, D., & Gloaguen, E. (2024). Quantifying uncertainty and improving prospectivity mapping in mineral belts using transfer learning and Random Forest: A case study of copper mineralization in the Superior Craton Province, Quebec, Canada. Ore Geology Reviews, 166, 105918 (16 pages). https://doi.org/10.1016/j.oregeorev.2024.105918

Titre de la revue: Journal Title:	Ore Geology Reviews (vol. 166)
Maison d'édition: Publisher:	Elsevier
URL officiel: Official URL:	https://doi.org/10.1016/j.oregeorev.2024.105918

Ore Geology Reviews 166 (2024) 105918

Contents lists available at ScienceDirect

Ore Geology Reviews

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

Quantifying uncertainty and improving prospectivity mapping in mineral belts using transfer learning and Random Forest: A case study of copper mineralization in the Superior Craton Province, Quebec, Canada

Dany Lauzon [a,b,*], Erwan Gloaguen [b]

[a] Civil, Geological and Mining Department, Polytechnique Montréal, C.P. 6079 Succ. Centre-ville, Montréal, Quebec H3C 3A7, Canada
[b] Centre Terre Eau Environnement, Institut National de la Recherche Scientifique, 490 Couronne St., Quebec City, Quebec G1K 9A9, Canada

ARTICLE INFO

Keywords:
Mineral prospectivity mapping
Transfer learning
Quantification of uncertainty
Data integration
Mineral belts in the Superior Craton Province

ABSTRACT

Mineral prospectivity mapping (MPM) involves identifying locations with a higher potential for mineral exploration based on a set of explanatory variables. In cases where there is a scarcity or absence of unfavorable sites that adequately represent the geological context for deposit discovery, generating synthetic negative data sets becomes necessary to employ a machine learning algorithm optimally. Moreover, when favorable sites are insufficient for deposit discovery within a geological zone, machine learning methods can potentially result in large and highly uncertain prospecting areas. This article proposed a concept based on transfer learning by applying the knowledge gained from mineral belt signatures in different geological zones to a related area. The positive training data were taken from five mineral belts distanced from each other, while the negative data were sampled using geological constraints based on the distance to occurrences and spatial associativity. The results demonstrate that transfer learning, combined with geological constraints applied to the creation of negative datasets, improves model performance and prediction of known deposits while significantly reducing uncertainties. Mineral prospectivity models for predicting potential copper formations were generated using data from the Quebec Government's spatial reference geomining information system, SIGEOM. The case study for this work focused on the geological province of the Superior Craton, which encompasses the vast majority of northeastern Quebec.

1. Introduction

Mineral Prospectivity Mapping (MPM) is a process that assimilates mining exploration data (i.e., exploratory variables) such as lithological, structural, and topographic maps, airborne magnetic, gravimetric, and radiometric surveys, and geochemical data to produce a map of the mineral potential to predict favorable areas for the discovery of new deposits. Approaches are generally based on either the geological and morphological knowledge of the deposit (knowledge-driven) or the geological signature extracted from the data (data-driven), or a hybrid approach (Carranza and Laborte, 2015). Knowledge-driven predictive mapping relies on experts in geology who have a deep understanding of geological formations and mineral indicators. This expertise enables them to identify and interpret new deposits (Daviran et al., 2022; Lusty et al., 2012). One can cite fuzzy logic (Boadi et al., 2022; Tao et al., 2021), index overlay (Boadi et al., 2022), Dempster-Shafer belief theory (Mohammadpour et al., 2021), analytical hierarchy process (Xi et al., 2023), and data envelopment analysis (Ali Hosseini and Abedi, 2015) as common approaches. In contrast, data-driven predictive mapping assigns evidential weights empirically based on the spatial relationship between the exploratory variables and known mineral deposits (Harris et al., 2001). Several data-driven methodologies can be recognized such as weights of evidence (Tao et al., 2021), artificial neural networks (Chen et al., 2022), Bayesian modeling (Mao et al., 2023), logistics regression (Zhao et al., 2023), support vector machine (Zhang et al., 2022), and decision tree analysis (Yin and Li, 2022).

The choice between knowledge-driven and data-driven approaches depends on the exploration context. Knowledge-driven techniques are applied in more "greenfields" exploration regions, i.e., regions with no or few known deposits (Daviran et al., 2022). Here, geologists rely on the predictive power of ore genesis models to find mineral deposits in previously unexplored areas or areas where they are not already known. On

* Corresponding author at: Civil, Geological and Mining Department, Polytechnique Montréal, C.P. 6079 Succ. Centre-ville, Montréal, Quebec H3C 3A7, Canada.
E-mail address: dany.lauzon@polymtl.ca (D. Lauzon).

https://doi.org/10.1016/j.oregeorev.2024.105918
Received 17 August 2023; Received in revised form 17 January 2024; Accepted 23 January 2024
Available online 28 January 2024

the other hand, data-driven modeling methods are useful in areas with known occurrences, like established mining camps, known as brownfields (Carranza and Laborte, 2015). In these areas, geologists search for deposits near or adjacent to an already operating mine, or well-documented areas. Alternatively, a combined approach may be practical, using data-driven techniques in areas with known deposits and applying the knowledge to areas without known deposits, assuming similarities in geological terrain and mineral exploration models (Harris et al., 2015).

The advent of machine learning has significantly heightened the utilization of classification methods in data-driven mineral prospectivity mapping. This method involves categorizing geographic locations into prospective or non-prospective areas, yielding results presented as binary maps or maps indicating a confidence range of prospectivity. Classification approaches are generally divided into two categories: (i) unsupervised methods and (ii) supervised methods. In unsupervised methods, data-driven techniques are employed to generate clusters based on inherent similarities and dissimilarities within the spatial dataset, providing a nuanced perspective on mineral prospectivity (Cheng et al., 2023; Esmaeiloghli et al., 2021; Jansson et al., 2022; Soltani and Imamalipour, 2022; Torppa et al., 2019). Conversely, supervised methods undergo a training process using specified training samples to enhance their predictive capabilities (Carranza and Laborte, 2015; Harris et al., 2022; Parsa et al., 2022; Rodriguez-Galiano et al., 2015; Silva dos Santos et al., 2022; Zuo and Carranza, 2011). Supervised machine learning methods require two data sets: 1- a class of interest representing the locations of known ore-formations (positive data) and 2- a contrast class such as locations without deposits (negative data) to train the model. Properly trained, these methods can predict a more precise delineation of prospective zones than manual data-driven methods by learning the relationships between the exploratory variables and the positive-negative data (Rodriguez-Galiano et al., 2015; Silva dos Santos et al., 2022; Zuo and Carranza, 2011). It is important to note that in ore-formation processes, the positive data, are rare non-random events, while negative data is the result of common and random geological processes in unknown locations (Silva dos Santos et al., 2022).

Previous studies have shown that randomly generated negative data sets can never be guaranteed to represent true negatives (Carranza, 2009). However, various techniques can be used to reduce the margin of error. One can use the spatial distribution of deposits to define areas where new deposits are less likely to be found (Carranza et al., 2008), while others may limit negative data to lithological units that are less favorable to ore formation (Nykänen et al., 2015), or a combination of both methods (Silva dos Santos et al., 2022). One can also choose negative data sets by considering the spatial locations of unrelated mineral deposits (i.e., mineral deposits that involve a commodity different from the commodity of interest). Such an approach introduces less uncertainty compared to the use of randomly selected negative datasets. (Lachaud et al., 2021). However, studies barely focus on the nature of positive data. Indeed, it cannot be certain that at a given site, positive data reflects the entirety of the ore-formation mechanisms or signatures for a given mineral. For example, copper deposits may be associated with hydrothermal fluids (Ross et al., 2020), massive volcanic sulfides (Mathieu, 2019), sedimentary exhalative deposits (Card and Poulsen, 1998), or nickel-copper (Ni-Cu) mafic–ultramafic complexes (Lawley et al., 2021), to name a few. Each of these mechanisms has a distinct geological signature that must be considered in the positive data to better predict the mineral potential into a given geological province using machine learning techniques.

One way to improve prospectivity mapping and reduce associated uncertainties can be by using transfer learning (TL). TL seeks to enhance comprehension of the current task by establishing connections with tasks conducted in different contexts but within a related source domain. This approach augments learning by establishing relationships between prior tasks and the target task, resulting in more logical, expedited, and better solutions (Hosna et al., 2022). In our application, TL addresses a research problem that involves leveraging knowledge acquired learning prospectivity indicators in well-known geological regions with multiple ore-formation signatures (i.e., brownfields), and applying it in an underexplored region or a region with limited positive examples (i.e., greenfields). For instance, the knowledge gained from learning to recognize mineralization in a well-documented geological sub-province can be applied to the task of identifying mineralization in a geologically distinct sub-province located far away from the original learning task.

In this article, a TL-based framework for quantifying uncertainty in mineral prospectivity mapping of critical and strategic minerals is proposed. Positive data are sampled across multiple geologic zones and settings to learn various deposit mechanisms, while negative data are generated at random locations constrained by spatial and geological criteria. Sparse geochemical data are integrated into the learning process using a regression interpolation method. The Random Forest (RF) algorithm was chosen for MPM because of its ability to randomly bootstrap and permute predictor maps throughout each tree. This randomness helps prevent the overfitting of the generated models. It is worth noting that RF has been effectively used for mineral prospectivity mapping by several authors, including Carranza and Laborte (2015), Harris et al. (2015), McKay and Harris (2016), Zhang et al. (2022); Zhang et al. (2019), Ford (2020), Parsa and Maghsoudi (2021), Silva dos Santos et al. (2022). The methodology is applied to a real case study for Cu prospectivity in the Superior Craton geological province, which covers most of the northeastern Quebec province in Canada.

2. Metallogenic and geological context of the Superior Craton Province

The Superior Province, located in Eastern Canada, encompasses a vast area of approximately 1.4 x 10^6 km^2, establishing itself as the largest Archean craton on Earth. Within this region, one can discover a diverse array of both common and rare rock formations, showcasing a wide range of metamorphic grades ranging from subgreenschist to granulite facies. This province serves as a host to numerous globally renowned mineral deposits, including exceptional reserves of gold and base metals (Ayer et al., 2008; Card, 1990; Card and Poulsen, 1998; Hathway et al., 2008; Percival et al., 2012) as well as many smaller but economically significant deposits.

The Superior Province contains a large variety of deposit associated with critical and strategic minerals that are classified based on their form, composition, and specific lithological associations (Card and Poulsen, 1998; Hathway et al., 2008). To name a few, deposits associated to volcanogenic massive sulphides (VMS), hydrothermal fluids deposits (HF), sedimentary exhalative deposits (SEDEX), Cu-Mo porphyries, Li-Cs-Be-Ta-Sn-U-Th pegmatites, Ni-Cu sulphides in layered gabbro/ultramafic sills, or Ni-Cu-PGE gabbro/ultramafic stocks are unique to the Superior Province. Those deposits are not uniformly distributed into the subprovinces, but rather are distributed in 'mineral belts' that contain deposits of diverse metal associations and of different types (Card and Poulsen, 1998).

Five regions containing mineral belts within the Superior Craton province were selected: I. Matagami region, II. Chibougamau area, III. Rouyn-Noranda District, IV. Troilus region, and V. North zone of the La Grande subprovince. These mineral belts are contained in three subprovinces: Abitibi, Opatica, and La Grande subprovinces (See Fig. 1). They were chosen for their various types of deposits associated with Cu mineralization, all mainly within greenstone belts (See Fig. 2). Table 1 summarizes the types of deposits found in each region where the occurrence of Cu acts as the principal commodity. All types of deposits have been utilized for the training and validation (see Table 1). We acknowledge that these deposits result from inherently diverse mechanisms. Nevertheless, accurately pinpointing these mechanisms constitutes a complex challenge. Our objective is to identify Cu commodity by leveraging exploratory variables, without making a distinction based on

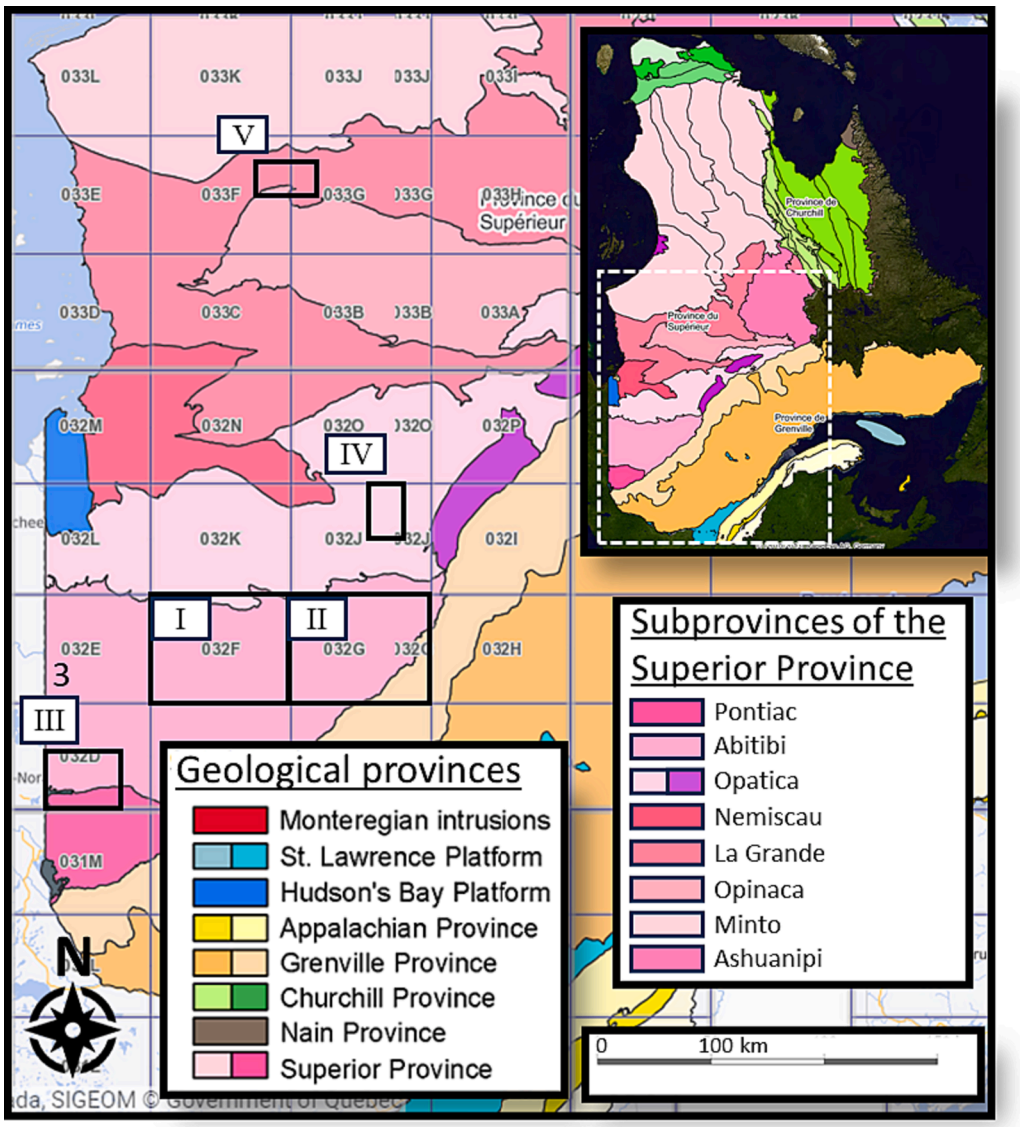

Fig. 1. Geological provinces of Quebec, Canada, and subprovinces of the Superior Province (pink units). The labeled black rectangles show the five study areas within the Superior Province. (I- Matagami region, II- Chibougamau area, III- Rouyn-Noranda District, IV- Troilus region, and V- North zone of the La Grande subprovince). Tags in the 000X format represent the NTS map-sheet number in the SIGEOM database, with NTS sheets delimited by gray lines. Figure adapted from SIGÉOM (2022). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the specific nature of the deposits.

The next subsections introduce a summary of the metallogenic and geological processes present in the selected areas: the Abitibi subprovince for areas I, II, and III; the Opatica subprovince for area IV; and the La Grande subprovince for area V (refer to Fig. 2 for labeling).

2.1. Abitibi subprovince

The Abitibi Subprovince is characterized by its composition, consisting of approximately 40 % metavolcanic and metasedimentary rocks primarily found in the Abitibi greenstone belt, and 60 % granitoid rocks, predominantly in the form of large batholiths. Within the greenstone belts, volcanic and related intrusive rocks account for around 80 % of the rock types, while metasedimentary rocks make up the remaining 20 %. The Abitibi greenstone belt encompasses numerous Neoarchean volcano-sedimentary sequences. It also comprises a diverse range of intrusions, varying from synvolcanic mafic-ultramafic complexes to post-tectonic granite. Significant volcanic activity occurred between 2.75 and 2.70 Ga, followed closely by deformation, regional metamorphism, and plutonism between 2.70 and 2.65 Ga (Card and Poulsen, 1998).

The Abitibi Subprovince hosts Au and Cu deposits which are primarily associated with HF, which are spatially corrolated with large-scale fault and shear zones (Harris et al., 2022). The subprovince is also enriched with base metal mineralization, particularly Cu and Zn, commonly found in association with VMS within the greenstone belts.

2.1.1. Matagami-Chibougamau mineral belt

The Matagami-Chibougamau mineral belt is of interest as it hosts mainly Zn-Cu VMS deposits, and Cu-Au HF deposits, and, of minor importance, intrusion hosted Ni-Cu deposits. The VMS deposits of the northern volcanic zone are hosted by mafic-felsic volcanic sequences that are older than 2.715 Ga. Vein Cu-Au deposits are hosted mainly by two distinctive layered intrusions of different ages, one of 2.728 Ga composed mainly of gabbro and anorthosite, and one aged of 2.716 Ga (Card and Poulsen, 1998). These deposits are mainly localized in shear zones that cut transversely across the layered intrusions.

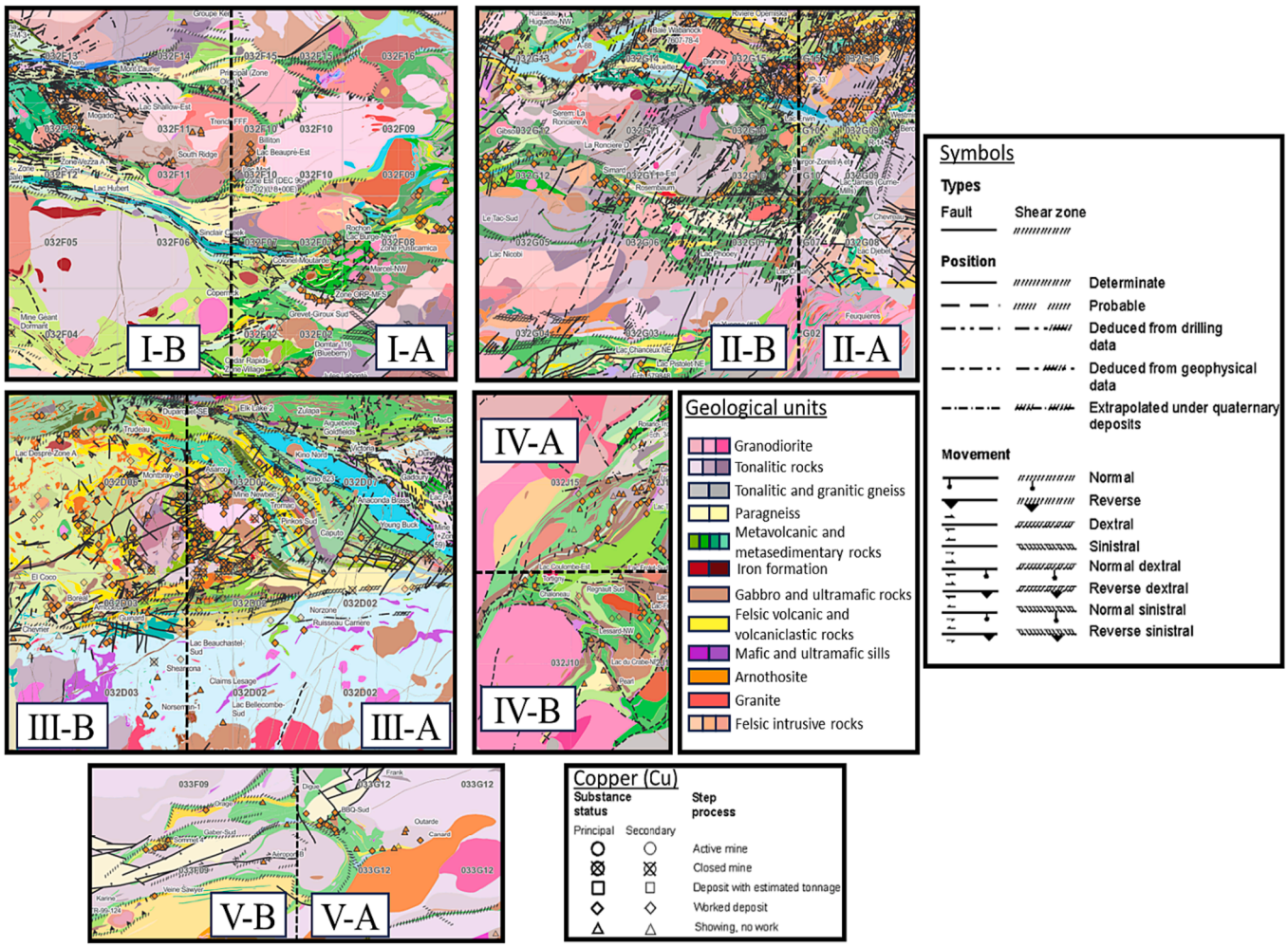

Fig. 2. Regional geological maps, fault locations (black lines), and copper occurrences (orange symbols) of the studied areas. Each area is divided into two (black dotted lines), one for training (zone A) and one for validation (zone B). Mineral belts are associated with metavolcanic and metasedimentary rocks, labeled in shades of green. The complete color codes can be consulted at the following address: MERN (2014). (I- Matagami region, II- Chibougamau area, III- Rouyn-Noranda District, IV- Troilus region, and V- North zone of the La Grande subprovince). Figure adapted from SIGÉOM (2022). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Number and proportions of each type of deposit (or occurrence) observed by region.

Region	VMS	HF	SEDEX	Ni-Cu	Others	Undefined	Total
1- Matagami	36 (32.43 %)	29 (26.13 %)	16 (14.41 %)	14 (12.61 %)	5 (4.51 %)	11 (9.91 %)	111
2- Chibougameau	26 (8.84 %)	180 (61.23 %)	29 (9.86 %)	7 (2.38 %)	31 (10.55 %)	21 (7.14 %)	294
3- Rouyn-Noranda	48 (17.91 %)	140 (52.24 %)	3 (1.12 %)	8 (2.99 %)	42 (15.67 %)	27 (10.07 %)	268
4- Troilus	26 (37.68 %)	23 (33.33 %)	5 (7.25 %)	1 (1.45 %)	7 (10.14 %)	7 (10.14 %)	69
5- La Grande	8 (16.67 %)	22 (45.83 %)	0 (0 %)	0 (0 %)	9 (18.75 %)	9 (18.75 %)	48

Note 1. VMS: Volcanogenic massive sulfide ore deposit, HF: hydrothermal fluid filings (copper and gold), SEDEX: Sedimentary exhalative deposits, Ni-Cu: Magmatic nickel-copper deposits, Others: porphyry deposits, veins of silver-lead-zinc, epithermal deposits, banded iron formation deposits, etc. Undefined: No information about the ore formation in the SIGEOM database.
Note 2. Copper mineralization is one of the main ore formations in the listed deposits.

2.1.2. Timmins-Val d'or mineral belt, Rouyn-Noranda district

The Timmins-Val d'Or mineral belt contains six main volcanogenic massive sulphide districts including Rouyn-Noranda District approximately aged of 2.70 Ga. The VMS deposits of the Rouyn-Noranda district are among the most intensively studied in the world. Most deposits may be linked to one major center of volcanism. The intrusions are inferred to be the subvolcanic heat sources that drove convective hydrothermal systems in overlying volcanic rocks that occupy a paleocauldron structure. Some deposits near the structure may have no direct relationship to the major volcanism episode but may occur in correlative (Card, 1990). Most of the deposits are from HF, or VMS with some others associated with SEDEX, and Ni-Cu ore deposits in mafic–ultramafic complexes (SIGÉOM, 2022).

2.2. Opatica Subprovince, Frotet-Evans-Troilus greestone belt

The magmatic-tectonic history of the Opatica Subprovince starts with an early period of high-grade metamorphism at ca. 2.718 – 2.721 Ga, and ends with strike-slip faulting occurring after 2686 Ma, with thermal activity occurring as late as 2.657 Ga. The regions have been

affected by several phases of deformation accompanied by low- to intermediate-pressure greenschist- and amphibolite-facies metamorphism. They are enclosed and intruded by voluminous granitoid rocks. In the northeastern part of the Archaean Frotet-Evans-Troilus greenstone belt lies the Troilus gold-copper deposit. The Troilus region contains many occurrences of Au, Cu, and Mo mineralization; Troilus being the largest Au-Cu deposit found to date.

2.3. La Grande subprovince

The La Grande subprovince is situated in the east-central region of the Superior Province in Québec Province. It is characterized by complex sequences of Archean volcano-sedimentary and plutonic rocks that have undergone multiple deformation events. The subprovince comprises a series of submarine volcanic sequences, including the Guyer Group (2.82–2.81 Ga) and the Yasinski Group (2.75–2.73 Ga), along with sedimentary sequences overlaying tonalitic gneiss basement (3.45–2.79 Ga). These rocks have been intruded by intermediate to felsic plutonic rocks (~2.72–2.71 Ga). The geological history of the La Grande area includes four significant compressional deformation events during the Archean period, accompanied by three associated metamorphic episodes. A prominent steeply dipping foliation, trending ENE (East-Northeast), dominates the region. The La Grande subprovince hosts a world-class gold deposit and several other deposits of base metals such as Cu, Zn, Ni, and Li (Mercier-Langevin et al., 2012; Sappin et al., 2018).

3. Materials and methods

MPM needs exploratory variables that reflect the characteristics of deposits. This section outlines the different steps required to obtain a mineral prospectivity map according to the proposed methodology. Firstly, the RF algorithm is presented. Then, explanatory variables representative of metallogenic deposit models are selected to generate predictive maps. In addition, subsampled data (here, geochemical data) are interpolated using an RF regressor. Finally, the sampling strategy for negative data is specified, and the methodology to address uncertainties is outlined.

The data is compiled from the SIGÉOM database (SIGÉOM, 2022) and the QGIS software as well as the Python programming language are used for data preparation, model training, and data modeling. The scikit-learn, SciPy, NumPy, pandas, shap, and Matplotlib libraries were used. All the scripts and associated data are available on GitHub in the Data Availability Statement section.

3.1. Random Forest

The Random Forest algorithm is a supervised learning method that is frequently employed for regression and classification tasks. It utilizes an ensemble of decision trees, where each tree is trained on a random subset of the training data. To enhance diversity within the ensemble, random feature selection is employed at each tree node (Breiman, 2001). In each decision tree, the internal nodes are split into two descendant nodes based on the value of one or more exploratory variables. If the exploratory variable is continuous, the split is determined by a cut-off value, while for categorical exploratory variables, a binary split is applied. The final prediction is obtained by combining the predictions of each optimized decision tree through averaging for regression tasks or voting, with the majority class selected as the final prediction for classification tasks.

The best predictor and the best split are obtained using a slicing criterion, a function that measures the quality of a split. For a regression task, one measure is the mean squared error (MSE), which is equal to variance reduction as feature selection criterion and minimizes the L2 loss using the mean of each terminal node. The mean squared error formula is given by:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

where n is the number of samples in the node, y represents the true target values of the samples in the node and $\hat{y}$ represents the predicted target values of the samples in the node. The split that results in the lowest MSE is selected as the best split during the training process of the decision tree in the Random Forest regressor.

For classification task, the algorithm aims, as an example, to minimize Gini impurity, a metric used to estimate the probability of incorrect classification at a specific node. The Gini impurity formula is given by:

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{n} p_i^2 \quad (2)$$

where p_i represents the probability of an instance belonging to class i within the node. The Gini impurity measures the degree of impurity or randomness in a node. It quantifies the likelihood of misclassifying a randomly selected instance in the node based on the class distribution. A lower Gini impurity indicates a purer node with more homogeneous class labels. By selecting splits that result in the greatest reduction in Gini impurity, the algorithm achieves a more homogeneous distribution of class labels, leading to improved separability and more accurate predictions (Breiman, 2001; Rodriguez-Galiano et al., 2014).

The optimal hyperparameters for the Random Forest algorithm were determined using a grid search strategy. It conducts a random search within a parameter space defined by the user, employing cross-validation on batches of the training set. A 5-fold cross-validation strategy with 100 iterations was implemented to explore the optimal values of RF-related hyperparameters. Table 2 shows the parameter space used in this study.

3.2. Transfer learning

The idea behind TL is to focus on applying the knowledge gained from solving a given task to a related one (Hosna et al., 2022). The TL model is applied in this study to learn the associations between Cu mineralization and the corresponding exploratory variables within four well-identified mineral fields. The trained model is then applied to improve the MPM within a poorly documented distant field. This methodology will be compared with a local approach (LA), which predicts prospectivity maps solely based on information from the surrounding area. LA partitions a specified area into two segments (e.g., I-A for training and I-B for validation, as defined in Fig. 2). In contrast, TL employs other regions (e.g., regions II, III, IV, and V, as defined in Fig. 2) as training sets to capture patterns and relationships from a more extensive set of input features and geological signatures related to mineral belts. It's noteworthy that TL is also validated using the same validation sets as the LA (i.e., region I-B). Note that set I-A is not used for training in TL to avoid spatial bias when comparing both approaches. Fig. 3 shows the difference between LA and TL.

To conduct both LA and TL approaches, we employed Random Forest (RF) to classify nodes as prospective areas or not. As mentioned previously, LA focuses the training on a specific area using data from the

Table 2
Set of hyperparameters of the Random Forest algorithm for the random search strategy.

Hyperparameter	Set
bootstrap	[*False*, *True*]
max_depth	[2, 5, 10, 20, 30, 40, 50, *None*]
max_features	[*auto*, *sqrt*]
min_samples_leaf	[1, 2, 3, 4, 5]
min_samples_split	[2, 3, 5, 10]
n_estimators	[100, 200, 300, 400, 500, 600, 800, 1000]

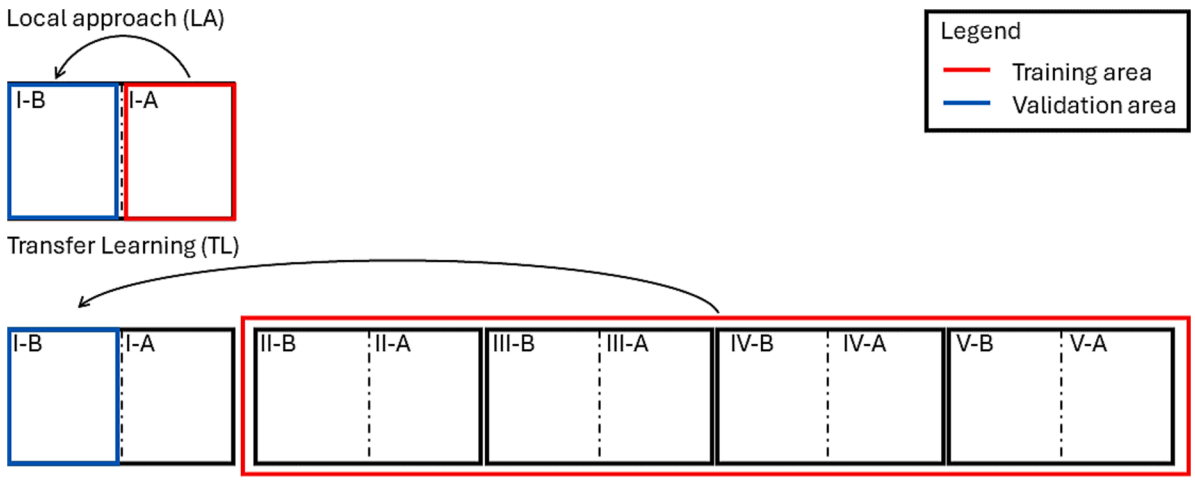

Fig. 3. The disparity lies in the training area between the local approach (LA) and the transfer learning methodology (TL). Data within the red area are employed for training, with validation performed on data within the blue area. In TL, four remote regions with a similar geological context are utilized to predict prospectivity, while in LA, the nearby area is used for predicting prospectivity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

vicinity of the validation zone. In contrast, TL utilized training data from various geological zones with similar regional metallogenic processes—specifically, greenstone belts with Cu mineralization. It can be inferred that LA is applied in a brownfield context, where we search for deposits near or adjacent to already operating mines. On the other hand, TL combines approaches by using data-driven techniques in well-known deposit areas (i.e., brownfields) and applying that knowledge to areas without known deposits or with only a few occurrences (i.e., greenfields). This assumes similarities in geological terrain and mineral exploration models, extending the predictive capabilities to promising regions where exploratory variables are available but lack positive data.

3.3. Preparation of exploratory variables

The exploratory variables are selected based on the formation mechanisms of the deposits present at the study sites and exploration data available from the SIGEOM database. These correspond to regional geological maps, fault and contact location maps, airborne magnetic and gravimetric geophysical surveys, and point geochemical analysis data. To simplify the reading, only the maps associated to the Rouyn-Noranda district, areas III-A and III-B, will be presented in this section. Note that all the features for all areas were re-interpolated or estimated to an identical regular grid with the same cell, a regular grid of 0.002° longitude by 0.002° latitude.

3.3.1. Indicator of regional geology

The regional-scale geological mapping of Quebec is the outcome of compiling works primarily conducted by government agencies (both provincial and federal) as well as mining exploration companies. It incorporates the results of numerous surveys, compilations, and geological syntheses conducted across the Quebec territory between 2012 and 2021. Additionally, high-resolution magnetic surveys conducted since 2012 have contributed to refining the geological interpretation. These extensive efforts, carried out at various scales, have facilitated the creation of relatively comprehensive geological coverage at the regional level. The Ministry of Natural Resources and Forestry (MRNF, 2023) produced the map at a scale of 1/2,000,000. The geology of the Rouyn-Noranda district is depicted in Fig. 2, area III.

3.3.2. Indicators of geological contacts and faults

Geological contacts and faults are known to be potential major conduits for fluid transport and sources of hydrothermal fluids (Harris et al., 2022). This information is represented by buffering geological contacts and fault locations from the SIGEOM database, creating two binary maps indicating the presence of contacts and faults. A buffer of 100 m was applied.

Subsequently, for each feature, two additional maps are generated. One indicates the distance to the nearest geological contact (or fault), and the other indicates the density of geological contacts (or faults). A high density suggests a higher likelihood of having a connected pathway for the circulation of hydrothermal fluids (Harris et al., 2022; Mathieu, 2019). The density is calculated using a moving average with a buffer of 500 m. The buffer was derived from the study by Harris et al. (2022), where the authors indicated that the influence of alteration and mineralization around faults and contacts extends around 500 m. The value was determined by the first author's experience while mapping various greenstone belts in Ontario, adjacent to the Québec Province, where greenstone belts are either the same or geologically similar. Fig. 3 and Fig. 4 show respectively the geological contacts and faults maps used for the Rouyn-Noranda district (Fig. 2, area III).

3.3.3. Indicators of geophysics

The compilation of low-resolution gravimetric data was carried out by the Geological Survey of Canada (GSC, 2023). All acquired data are referenced to the IGSN71 reference level to create a consistent dataset on a planetary scale. The data used to create the maps are interpolated on a 2 km grid with an interpolation limit of 20 km, including data acquired from 1944 to 2015. Currently, most base stations have been replaced to achieve a precision of 2 μGal (2x10-8 m/s^2). The Bouguer anomaly was calculated using a vertical gravity gradient corrected for free-air with a value of 0.3086 mGal/m and a density of 2670 kg/m^3 for Bouguer correction. The maps used for the prospectivity mapping are the Bouguer anomaly and the first vertical derivative map of gravity anomalies. Fig. 5 shows the normalized low-resolution gravimetric maps used for the Rouyn-Noranda district (Fig. 2, area III).

The compilation of high-resolution magnetic data was carried out by the Ministry of Natural Resources and Forests, Québec (MRNF, 2023). The surveys used for this map have a line spacing ranging from 100 to 300 m with a flight height between 40 and 100 m. The values of the residual component of the total magnetic field are interpolated on a 75 m grid. The map of the residual magnetic field of Quebec was created using the GridKnit tool from Geosoft. The first vertical derivative and the analytic signal were calculated using the Fourier transform method with the MAGMAP module from Geosoft. The maps used for the prospectivity mapping are the residual magnetic field, the first vertical derivative, and the analytic signal. Fig. 6 shows the normalized high-resolution

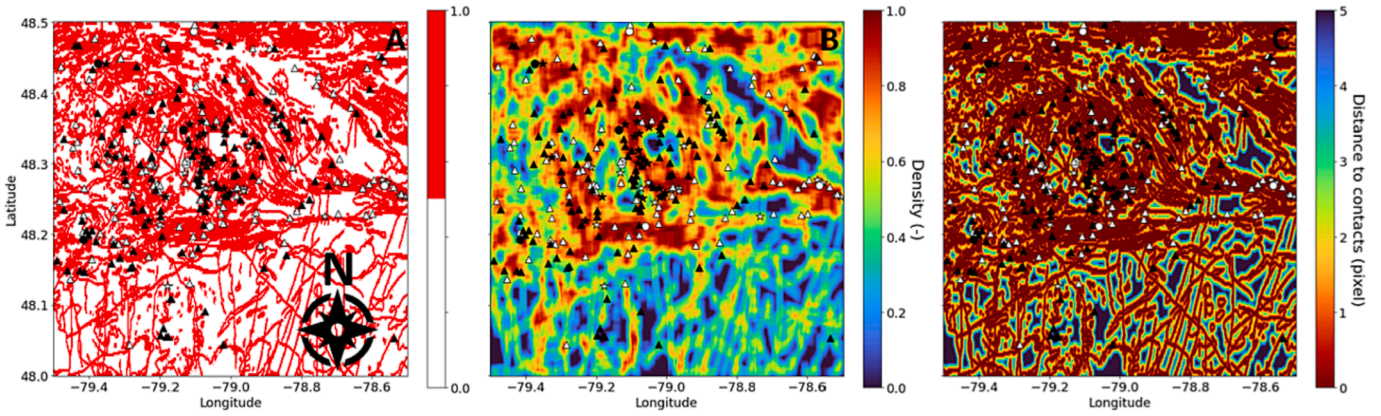

Fig. 4. Features associated with geological contacts. A – Binary geological contact location map, B – Geological contact density prediction map. C – Geological contacts distance map. (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, stars: mines, circles: deposits, Triangle: occurrences).

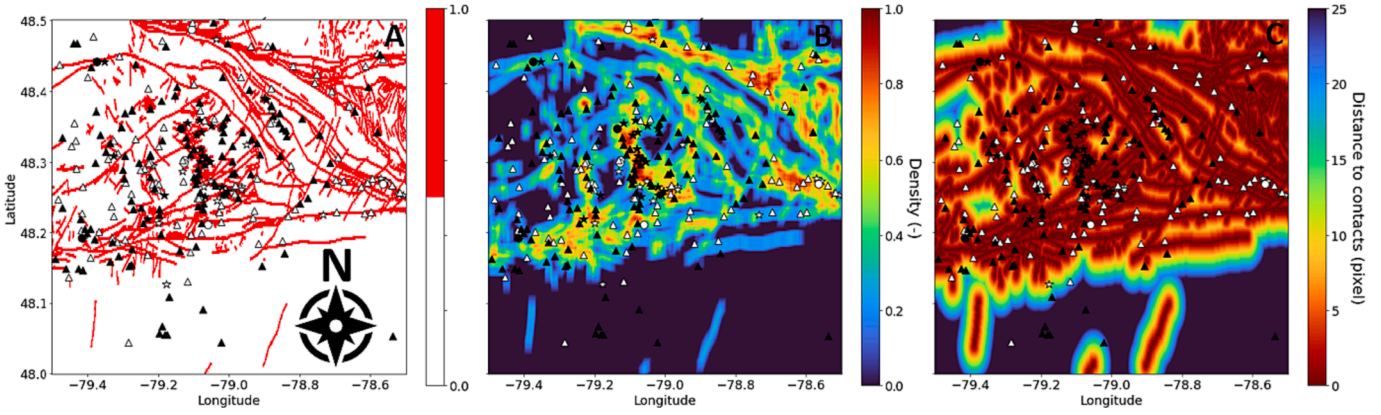

Fig. 5. Features associated with faults. A – Binary fault location map, B – Fault density prediction map. C – Fault distance map. (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, stars: mines, circles: deposits, Triangle: occurrences).

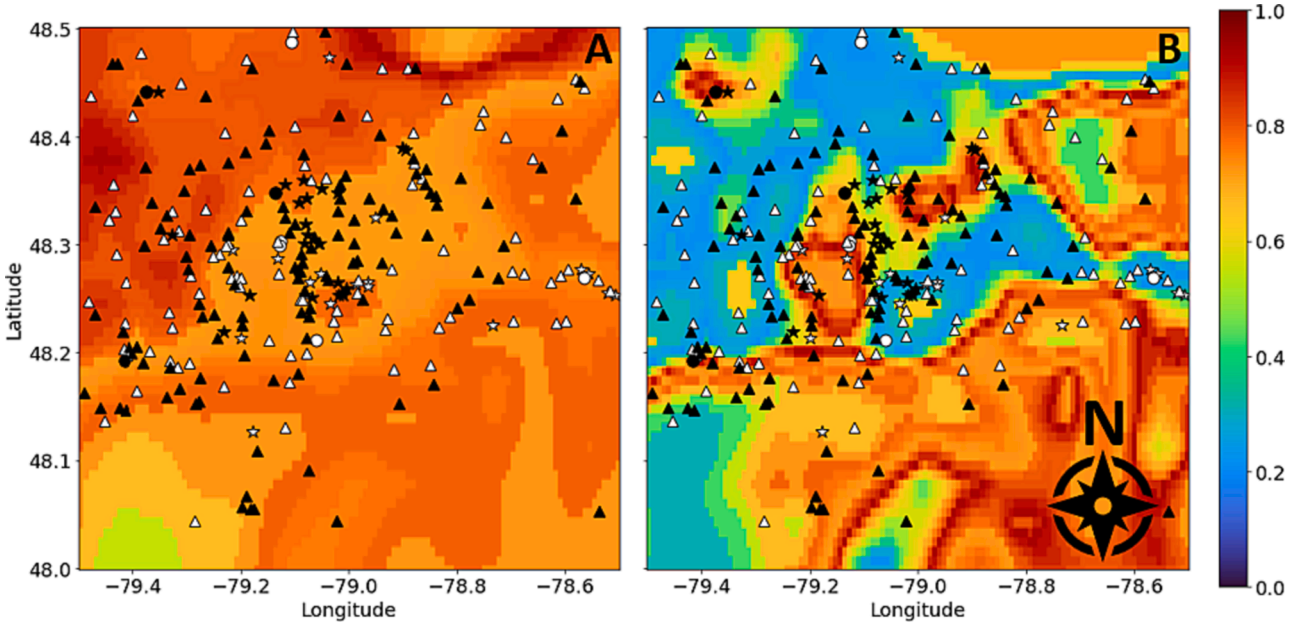

Fig. 6. Features associated with low-resolution gravity data. A – Map of the Bouguer anomaly, B – Map of the first vertical derivative of gravity anomalies. (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, stars: mines, circles: deposits, Triangle: occurrences).

magnetic maps used for the Rouyn-Noranda district (Fig. 2, area III).

3.3.4. Indicators of geochemistry

Geochemical data, which offer insights into rock types, alteration minerals, and anomalous concentrations, are of immense value in mineral exploration (Lachaud et al., 2023). In our study, only the chemistry of rocks has been considered. The data was obtained from SIGEOM and processed according to the methodology presented in the following section. These indicator elements, Au, Pb, Ni, Zn, Cu, were used in the modeling process, along with ten oxides (K_2O, MgO, Na_2O, CaO, Fe_2O_3, MnO, SiO_2, Al_2O_3, P_2O_5, TiO_2). Note that geological processes, such as differentiation, alteration, mineralization, and weathering, can be characterized by geochemical alteration indexes. These indexes serve as valuable tools for understanding and quantifying the intensity or extent of alteration zones, a mechanism present in our study (Lachaud et al., 2023; Mathieu, 2018). These indexes, derived from oxide percentages, include the carbonate-chlorite-pyrite index (CCPI) (Eq. (3)), the Ishikawa alteration index (IAI) (Eq. (4)), the advanced argillic alteration index (AAAI) (Eq. (5)), and the Altered-Unaltered index (UAI) (Eq. (6)). Both high values of IAI and CCPI indicate the presence of hydrothermal alteration in the altered products. An increase in AAAI represents strong SiO_2 enrichment and chlorite, carbonate, and feldspar destruction. An AUI value higher than 1 may indicate altered rocks, while unaltered rocks may have an AUI value lower than 1 (Mathieu, 2018). It is important to note that all these indexes represent different aspects of hydrothermal alteration and ore formation, reflecting the chemical and mineralogical changes that occur during the weathering process and indicating the density of alteration. For a comprehensive review of methods for alteration indexes (CCPI, IAI, UAI), refer to Mathieu (2018), and for insights into AAAI, consult Williams and Davidson (2004).

$$\mathrm{CCPI} = \frac{Fe_2O_3 + MgO}{Fe_2O_3 + MgO + Na_2O + K_2O} \tag{3}$$

$$\mathrm{IAI} = \frac{K_2O + MgO}{K_2O + MgO + Na_2O + CaO} \tag{4}$$

$$\mathrm{AAAI} = \frac{SiO_2}{SiO_2 + 10(MgO + Na_2O + CaO)} \tag{5}$$

$$\mathrm{AUI} = \frac{FeO + MnO + MgO}{CaO + K_2O + Na_2O} \tag{6}$$

Integrating geochemical data into the machine learning process requires transforming the point geochemical dataset into geochemical maps. Generally, the inverse distance method and Kriging-based methods are commonly used to interpolate geochemical data across the grid. Nevertheless, the results of such interpolations are generally smooth and not constrained to lithological units. To achieve this, RF regressor is employed, which allows evaluating the similarities between observation locations and geographical and geological attributes, namely exploratory variables, such as geological contacts, fault locations, geophysical data, regional geology, and any other parameters deemed relevant for estimation. By employing an RF regressor, our method fulfilled the requirements of our study, conditioning geochemical data accordingly to rock lithology and exploratory variables without the need for additional assumptions associated with more advanced techniques like regression kriging. All feature maps shown in section 3.3.1 to 3.3.3 were used as exploratory variables to predict geochemical data. Before conducting the interpolation, the geochemical data were averaged at the grid cell level to reduce noise. Grid data are not provided in the regressor to enable a more accurate estimation of geological units lacking measured geochemical data. Regression scores surpass 0.80 for all conducted regressions. Fig. 7 illustrates the rock geochemical maps for CCPI, AUI, SiO_2, Al_2O_3, Au, Cu, Zn, and Ni obtained for the Rouyn-Noranda district (Fig. 2, area III). The resulting maps demonstrate consistency with geology, as shown by the SiO_2 map, where the granitic units in the centers exhibit grades varying between 70 and 75 %, which are typical values for granite.

3.4. Sampling negative data

Accurate selection of training and validation sets is an important step in the Random Forest algorithm to learn patterns and relationships from the input features (i.e., exploratory variables) and target labels (i.e., deposits or non-deposits). However, in many applications in geosciences, these sets are biased by the inherent spatial correlation. These sets require to be sampled accordingly to the geological problem to prevent overfitting and to help detect when the model becomes too complex and fails to generalize well to new data (Silva dos Santos et al., 2022).

One proposed solution is to geographically divide the training and validation sets, such as allocating one in the northern region and the other in the southern region (Silva dos Santos et al., 2022). This is done by dividing area in sections A and B as shown in Fig. 2. The split between sections A and B is made to achieve a fair distribution of training and validation data across both sections. The idea is to create training and validation zones that are likely to be similar without being entirely identical. If they were identical, assessing TL gain over LA could be challenging, as the validation zone would closely resemble the training zone. Moreover, constraints on the negative datasets are necessary to sample signatures that are more likely to be inert, without

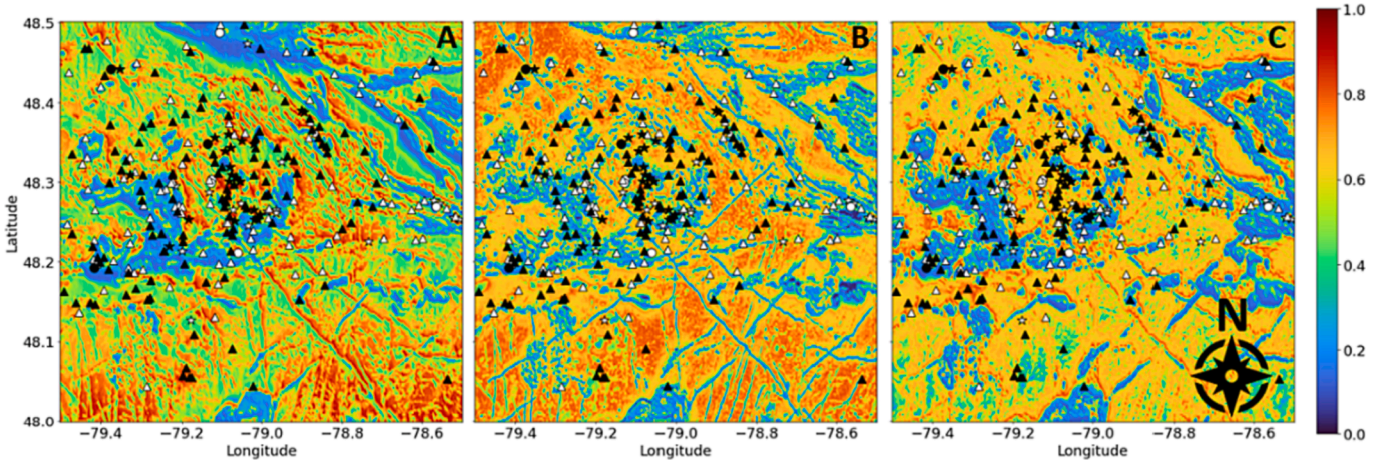

Fig. 7. Features associated with high-resolution magnetic data. A – Residual magnetic field map, B – Map of the first vertical derivative of the residual magnetic field, C – Analytical signal map. (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, stars: mines, circles: deposits, Triangle: occurrences).

mineralization. Two constraints can be applied to sample the negative data (Carranza et al., 2008; Nykänen et al., 2015; Silva dos Santos et al., 2022). The first constraint assesses the favourability of lithologic units to represent the strength of spatial association (SA) between rock units (R) and known deposits (D) (Nykänen et al., 2015; Silva dos Santos et al., 2022). SA, Eq. (7), is calculated by comparing the proportion of a unit's area covered by known deposits, P(R∩D), to the proportion of the study area occupied by the same unit, P(R), as:

$$SA(R) = \frac{P(R \cap D)}{P(R)} = \frac{N(R \cap D)/N(D)}{N(R)/N(A)} \quad (7)$$

where $N(R \cap D)$ represents the number of known deposits of type D contained in the rock unit R, N(D) is the number of known deposits of type D, N(R) is the number of pixels associated with the rock unit R, and N(A) the number of pixels of the study area A. To generate the spatial association map, we assign the SA value calculated for the rock unit present at the coordinates. Fig. 8A presents SA of region III. This metric helps prioritize negative samples from lithologic units that show less association with known deposits. The second constraint involves point spatial analysis to determine the proximity of deposits to each other in space (see Fig. 8B). This analysis helps identify how closely the deposits are located, aiding in the selection of negative samples that are spatially distant from known deposits (Carranza et al., 2008). Negative data are drawn randomly from areas with SA below the 25 % quantile. This enables the sampling of geological units with a significantly low probability of containing an occurrence due to the scarcity of observed occurrences within these units and their extensive surface area. Fig. 8 shows the applied constraints to the Rouyn-Noranda district (Fig. 2, area III).

3.5. Assessing uncertainties

The uncertainty of the predictions was assessed by generating 200 sets of negative data. For each set, the negative dataset was combined with the positive dataset from the study area. The combined dataset was then divided into training and testing sets based on sections A and B defined in Fig. 2. The RF model was trained using the training set, and subsequently, a potential map of the test area was generated using the trained model. The process was repeated 200 times to capture variability, and predictions and metrics were stored for each realization. Finally, the mean and standard deviation of the potential maps were calculated.

4. Results and discussion

The study is divided into two sections to highlight the advantages of TL in the mineral belt for Cu prospectivity mapping. It is worth mentioning that the methodology can be applied to other critical and strategic minerals such as zinc, nickel, or lithium, for instance. First, the well-documented Rouyn-Noranda area (Section III-B, Fig. 2) will be utilized to validate the approach by comparing the traditional LA with the more comprehensive regional approach based on TL. Second, the La Grande region (Section V-B, Fig. 2) will showcase the advantages of the proposed approach in cases where the surveyed region is not thoroughly explored. To lighten the results section, we focused on presenting results for zones III and V—a well-documented region (III) and a sparsely documented remote area (V). Although studies were conducted for regions I, II, and IV, the results closely mirror those of regions III and V. Fig. 1 and Fig. 2 visually depict the training and validation zones for these regions, but their results will not be presented.

Table 3 summarizes the number of mines, deposits, and occurrences found in each area where Cu is the main commodity. Field sizes are also indicated. Note that sets A and B have been created based on Fig. 2 by converting the coordinates of Cu deposits into a regular grid of 0.002° longitude by 0.002° latitude. If several ore formations are in the same grid cell, the most important formation (mine followed by deposits and occurrences) is selected.

4.1. Transfer learning approach on a well-documented dataset

The first study site, the Rouyn-Noranda district, is one of the most extensively studied VMS ore deposits in the world. Mineralization is well documented and geochemical analyses are present in large quantities (i.e., over 35,000 analyses). The intrusions are inferred to be the subvolcanic heat sources that drove convective hydrothermal systems in overlying volcanic rocks that occupy a paleocauldron structure (Card, 1990).

The training and validation set for LA separate the paleocauldron

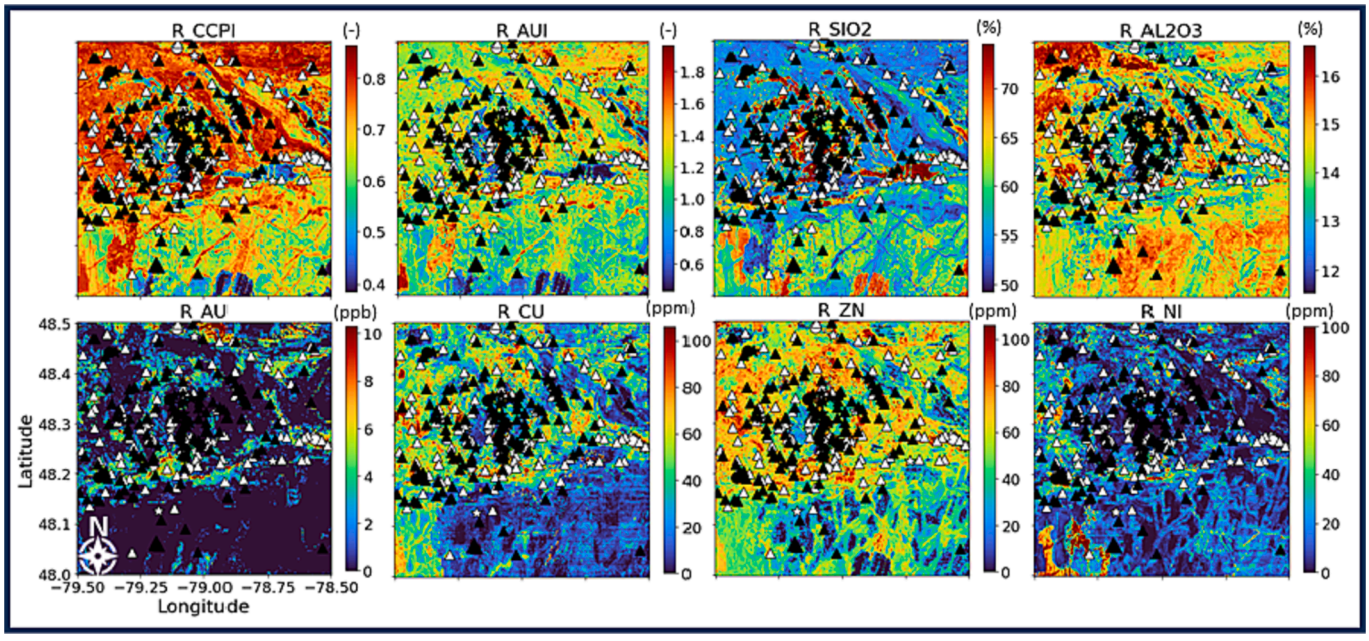

Fig. 8. Features associated with geochemical data. (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, stars: mines, circles: deposits, Triangle: occurrences).

Table 3
Number of mines, deposits, and occurrences per region and per section, and field dimensions.

	Section A			Section B			Dimensions	
Region	Mine	Deposit	Occurrence	Mine	Deposit	Occurrence	Long.	Lat.
I- Matagami	–	3	49	13	5	38	2°	1°
II- Chibougameau	18	3	112	6	4	150	2°	1°
III- Rouyn-Noranda	18	–	62	5	3	59	1°	0.5°
IV- Troilus	–	–	32	–	2	19	0.5°	0.5°
V- La Grande	–	–	20	–	–	18	1°	0.25°

structure into two parts consisting of regions III-A for training and III-B for validation (refer to Fig. 2). In contrast, TL uses the other eight regions (I-A, I-B, II-A, II-B, IV-A, IV-B, V-A, and V-B) as training sets to learn patterns and relationships from a larger number of input features and mineral belt geological signatures. Note that III-A is removed from the training sets to avoid biased from the inherent spatial structure of the exploratory variables of region III.

4.1.1. Mineral prospectivity mapping and uncertainty quantification

For each of the two approaches mentioned above, two hundred potential maps were generated. Subsequently, mineral prospectivity maps are produced, including an average predicted prospects map and a mean standard deviation map indicating prediction inconsistency in test areas (Fig. 9). The Maps in the left and right columns of Fig. 9 relate to the LA and the TL approach, respectively.

Maps generated using TL exhibit more promising results compared to the LA. This is evident through an increase in mean probabilities, and a decrease in mean standard deviations. Quantitatively, the probabilities associated with occurrences increased from an average of 73 % to 79 %, and the mean standard deviation decreased from 6 % to 5 %. Also, TL significantly enhances the prediction of hotspots, with an average probability of 94 % for mines compared to 81 % in the LA, and better predict secondary Cu mineralization (white data, not provided in the training sets) than the LA with the probabilities associated with occurrences increased from an average of 75 % to 83 %, and the average standard deviation slightly decreases from 5 % to 4 %. Table 4 provides a summary of these statistics categorized by type of mineralization, i.e., mine, deposit, or occurrence. Overall, prospectivity zones seems better defined (i.e., >50 %): adopting the LA, the 200 models predicted on average probabilities of 67 % with mean standard deviation of 6 %;

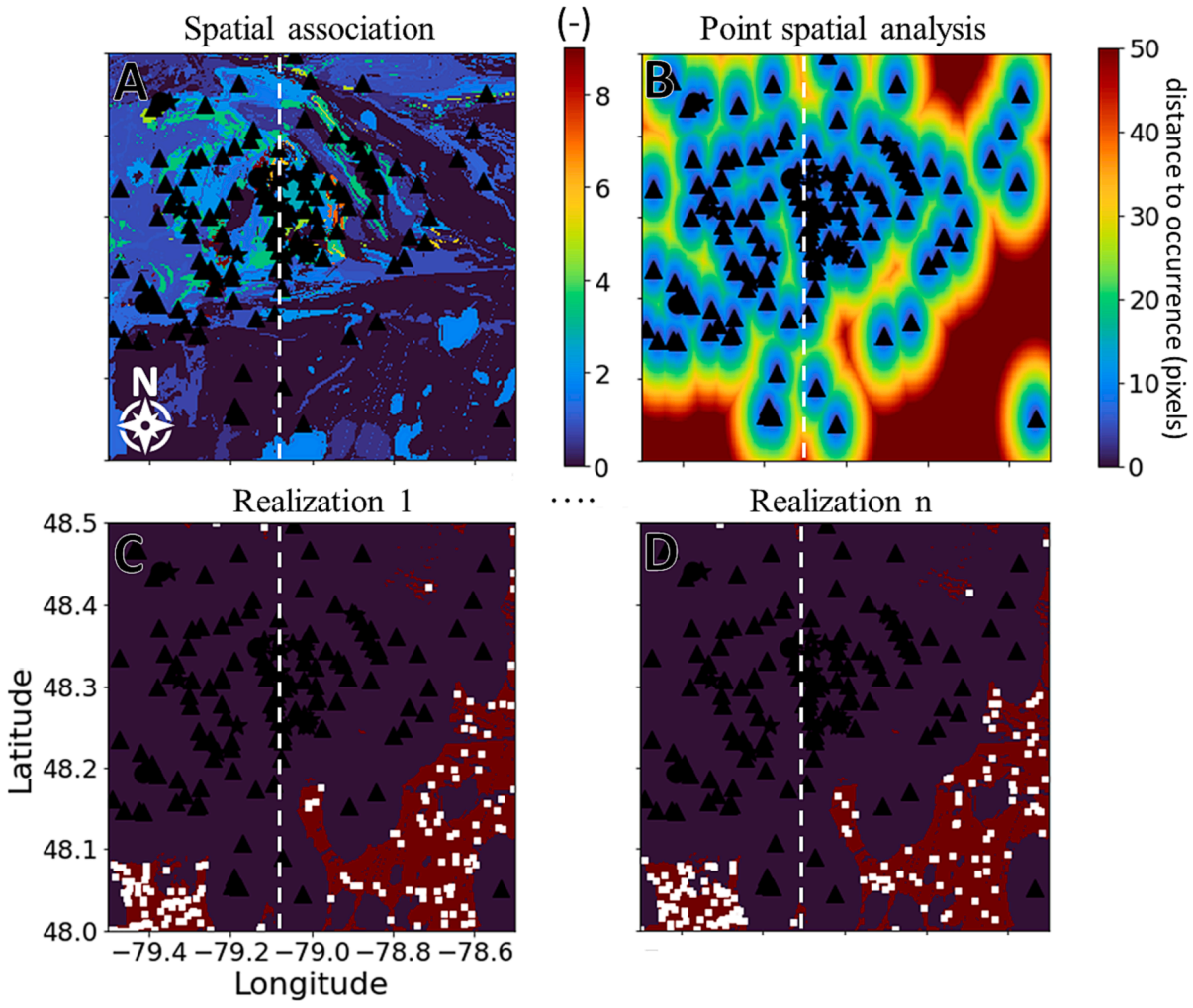

Fig. 9. Spatial association map (A). Point spatial analysis map (B). Realization 1 of negative data (C). Realization 95 of negative data (D). (Symbols – black: main Cu mineralization, stars: mines, circles: deposits, Triangle: occurrences, white squares: negative data).

Table 4
Summary of average probability and standard deviation by type of mineralization, i.e., mine, deposit, or occurrence for local approach (LA) and transfer learning (TL) for primary (PM) and secondary (SM) mineralization.

	Average probability (%)			Average standard deviation (%)		
	Mine	Deposit	Occurrence	Mine	Deposit	Occurrence
PM- LA	81	76	66	5	6	6
PM- TL	94	82	71	3	4	5
SM- LA	75	77	68	5	6	6
SM- TL	83	79	72	4	5	5

while adopting the TL approach, the average probabilities become 73 % with mean standard deviation of 5 %.

4.1.2. Impact of TL on prediction of hotspots for exploration targeting

Two categorical maps were generated to highlight prospecting areas. Fig. 10(A and B) illustrates the region where the average probability exceeds 50 %, indicating promising areas. Fig. 10(C and D) incorporates thresholds from Fig. 9(A and B) and assigns labels based on average probabilities. Areas with probabilities above 90 % are labeled red, those between 80 % and 90 % are labeled gold, between 70 % and 80 % are labeled yellow, between 60 % and 70 % are labeled lime, and between 50 % and 60 % are labeled green. Take note of the black circles in Fig. 10D, which indicate high-potential areas lacking known Cu occurrences. These areas appear to be promising candidates for future exploration targets.

Firstly, there is a significant decrease in the prospecting area passing from 63.30 % to 57.91 %. These areas are mainly associated with intrusive igneous rocks and have a low probability of containing copper mineralization. By incorporating the four other mineral belts, the model provides a better characterization of the inert signature of intrusive igneous rocks, allowing us to narrow down the search areas to the most attractive ones. Moreover, TL better identifies the hotspots for exploration targeting compared to the LA. This phenomenon is shown in the increase in the number of principal occurrences in the red-gold categories, from 21 to 39, or from 18 to 25 for secondary mineralization.

4.1.3. Feature importance

Silva dos Santos et al. (2022) emphasize the importance of reliable prospecting maps, which necessitate models built upon reasonable input data, including proxies that represent various mechanisms responsible for ore formation. However, achieving reliable and interpretable results is equally crucial, and this can be challenging due to the inherent "black box" nature of machine learning models. To shed light on this issue, SHapley Additive exPlanations (SHAP) values are computed to provide explanations for individual model predictions. SHAP is a method derived from cooperative game theory used to attribute a value to each feature in a prediction (Lundberg and Lee, 2017). They aim to explain the output of machine learning models by quantifying the contribution of each feature to the model's prediction for a specific instance. SHAP values provide a way to distribute the "credit" for a model's prediction among its input features, offering insights into the importance and impact of each feature on the final output. Fig. 11 displays the SHAP values associated with the 200 trained models using the LA (A) and TL (B). All simulated values are incorporated in Fig. 11.

The LA consistently places significant emphasis on gravity surveys (BouguerBR and AnoGrav1DBR), as well as fault characteristics (Fault_Distance and Fault_Density). The algorithm suggests that mineralization is predominantly linked to medium-low density rocks, such as those found in the greenstone belt, and highlights the importance of fault locations, indicating the probable influence of hydrothermal fluids on Cu-mineralization. On the other hand, the TL approach presents a different perspective. It focuses more on structural geology, one of the

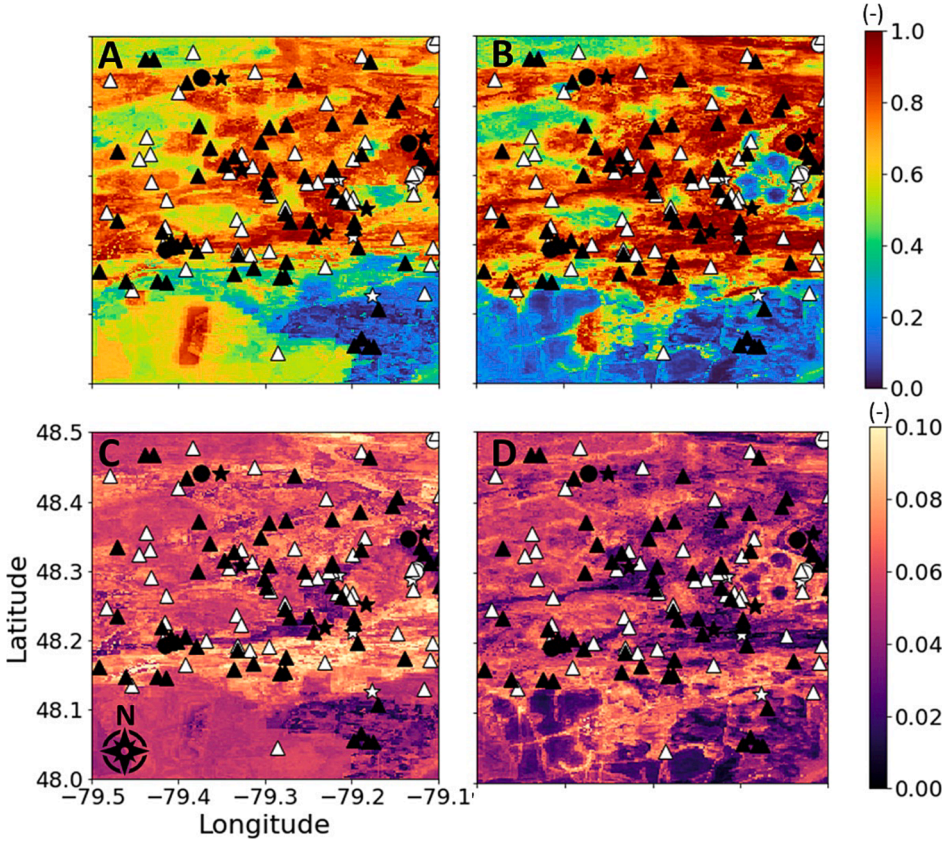

Fig. 10. Potential maps (A-B). Standard deviation maps (C-D). Left column: local approach (LA). Right column: transfer learning (TL). (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, stars: mines, circles: deposits, Triangle: occurrences).

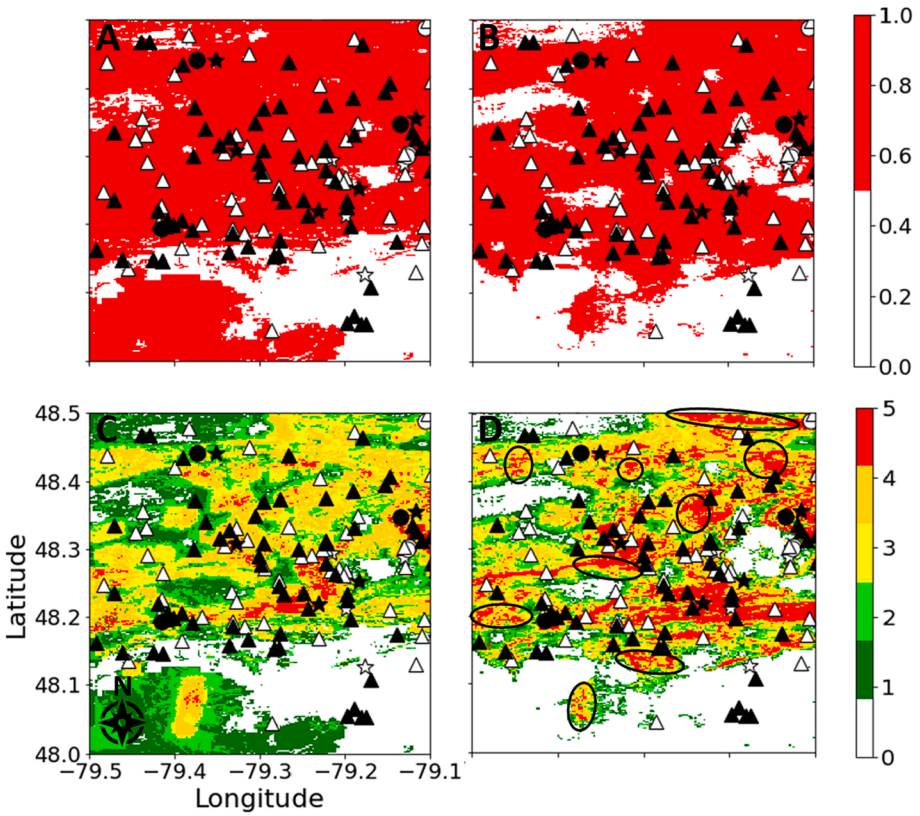

Fig. 11. Prospecting maps (A-B). Labeled prospecting maps (C-D). Left column: local approach (LA). Right column: transfer learning (TL). (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, stars: mines, circles: deposits, triangle: occurrences).

main migration mechanisms for hydrothermal fluids, with a strong emphasis on fault locations and geological contacts. In addition, importance is given to rocks with high CCPI (R_CCPI) and AUI (R_AUI) indices, indicating locations where rock units have potentially been altered because of processes such as chloritization, an alteration of magmatic rocks characterized by hydration and potassium release, or the passage of hydrothermal fluids (Mathieu, 2019). This disparity in emphasis illustrates that the geological interpretation of the results differs between the two methods. The TL method seems to capture a greater geological signature of the underlying mechanisms associated to Cu commodity.

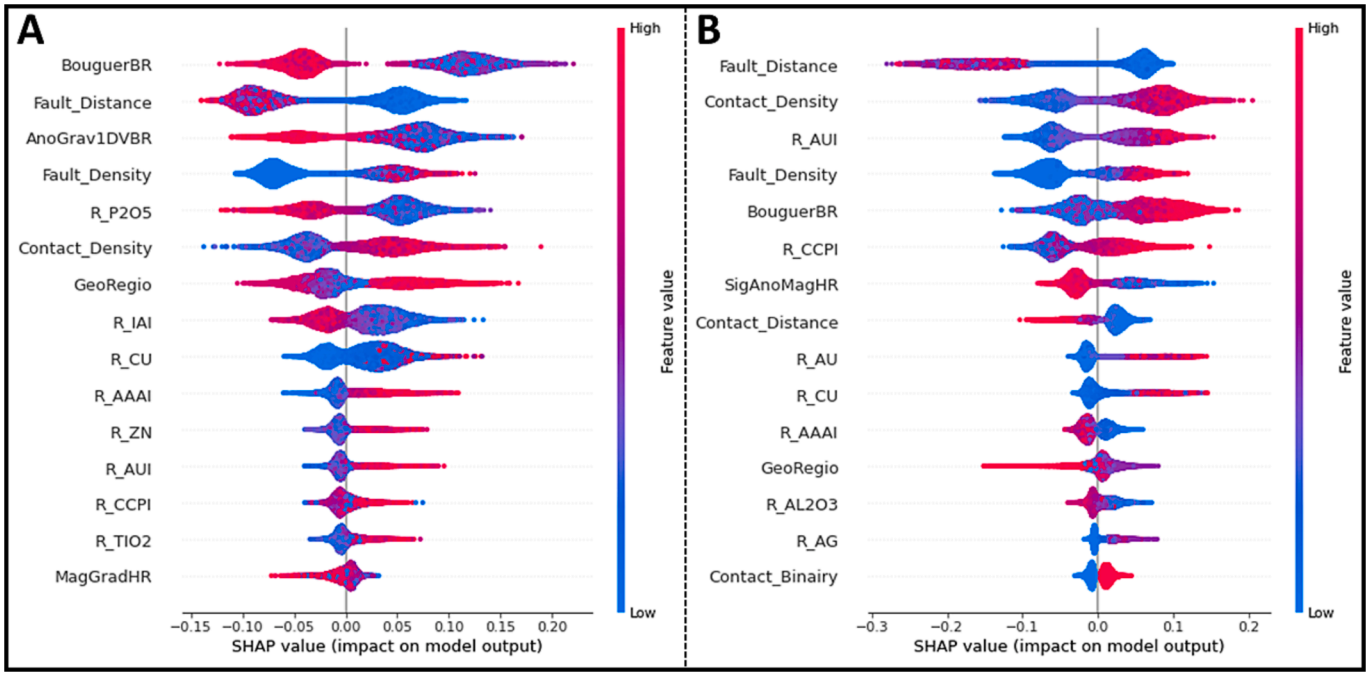

Fig. 12. SHAP value of the feature importance. A) Local approach (LA). B) Transfer learning (TL).

4.2. Transfer learning approach on a remote region

The La Grande sub-province is located within the central-eastern part of the Superior Province of Quebec (See Fig. 1, area V). It is characterized by complex sequences of Archean volcano-sedimentary and plutonic rocks that have undergone multiple deformation events. The subprovince hosts a world-class gold deposit and several other base metal deposits such as Cu, Zn, Ni and Li (Sappin et al., 2018). These recent discoveries have increased interest in the region, underscoring the need to better define areas of mineral prospectivity. Exploration and new regional mapping provide an improved geological framework, but some areas lack positive data to capture the geological signature of Cu mineralization by machine learning. TL should therefore be favored.

The training and validation set for the LA separate the zone into two equal parts consisting of regions V-A for training and V-B for validation (refer to Fig. 2). In contrast, TL uses the eight following regions I-A, I-B, II-A, II-B, III-A, III-B, IV-A, and IV-B as training sets. For each of the two approaches, two hundred potential maps were generated. Subsequently, mineral prospectivity maps are produced, including an average predicted prospects map, a mean standard deviation map indicating prediction inconsistency in test areas, and a hotspots map for exploration targeting (Fig. 12). The left column represents the LA, and the right column is associated with the TL approach. Take note of the black circles in Fig. 12D, which indicate high-potential areas lacking known Cu occurrences. These areas appear to be promising candidates for future exploration targets. Fig. 13.

Maps generated using TL yield more promising results compared to the LA. This is evident through higher average probabilities, lower average standard deviations, and improved identification of areas of interest for exploration. In terms of quantitative analysis, the probabilities associated with occurrences have increased from an average of 81 % to 89 %, while the average standard deviation has decreased from 5 % to 4 %. Overall, the hotspots areas appear to be more precisely delineated and closely aligned with observed occurrences. For instance, using

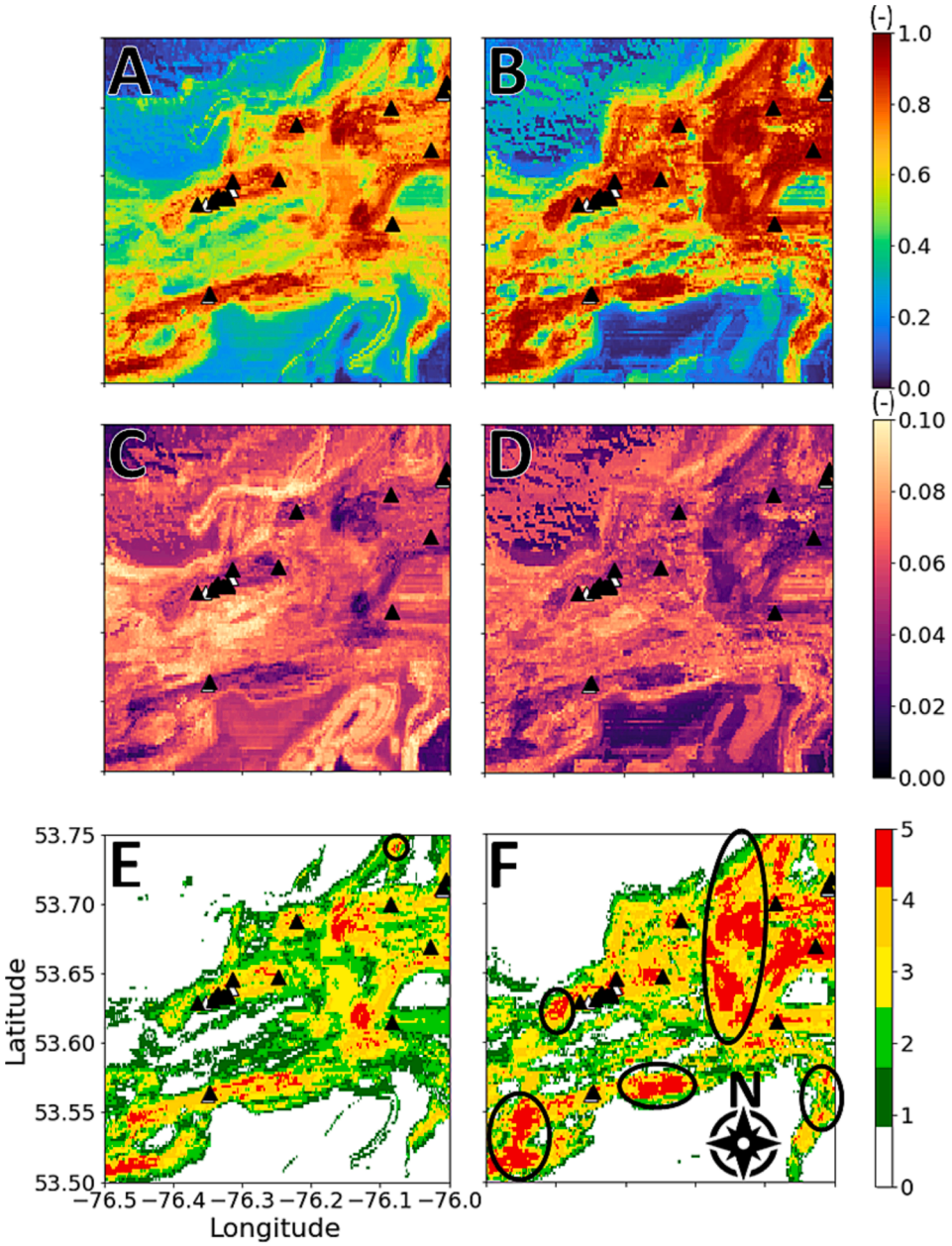

Fig. 13. Potential maps (A-B). Standard deviation maps (C-D). Labeled prospecting maps (E-F). Left column: local approach (LA). Right column: transfer learning (TL). (Symbols – black: main Cu mineralization, white: secondary Cu mineralization, Triangle: occurrences).

the LA, the 200 models predicted mean probabilities of 70 %, whereas the TL achieved mean probabilities of 75 %. This is also demonstrated by an increase in the number of occurrences in the red-gold category, rising from 12 to 22 (including secondary mineralization). It is important to note that the TL indicates no occurrence out of 22 in the green, lime, and yellow categories, whereas the local method identifies 12 out of 22, corresponding to 55 % of the occurrences. This significant difference can be attributed to the limited number of positive data points from region V, amounting to 38 occurrences for LA as opposed to 599 for TL. The difference between the number of mines/deposits/occurrences in Table 1 (742 in total) and the 599 occurrences quoted here, is related to several mineralized formations present in the same grid cell. It appears that the TL provides a good understanding of Cu ore formation mechanism in greenstone belts and can be used to predict under-sampled or under-documented remote areas.

4.3. Impact of TL on seven different supervised machine learning methods

To highlight the potential of TL in mineral prospectivity mapping, we reconducted experiments in sections 4.1 and 4.2 using various supervised machine learning algorithms, including logistic regression (LR), k-nearest neighbors (KNN), support vector machine (SVM), artificial neural networks (ANN), adaptative boosting (AdaBoost), and extreme gradient boosting (XGBoost). The training and validation methodology is identical with that presented in section 4.1 for the well-documented region (brownfield) and in section 4.2 for the remote region (greenfield). Tables 5 and 6 present the validation results for regions III-B, the well-documented region, and V-B, the remote region, respectively. RF results are also provided.

Four metrics—Brier Score (BS), Recall (RC), Predicted Probability of Mineralization (P_{PM}), and the Prospecting Area (P_A)—were computed to compare LA to TL. BS provides a measure of the mean squared difference between predicted probabilities and actual outcomes, offering a way to evaluate the calibration and accuracy of probabilistic predictions. Lower BS indicates better predictive performance. RC measures the ability of a model to correctly identify all relevant instances (positives) from the total number of actual positive instances in a dataset. Lower RC indicates better predictive performance. P_{PM} is associated with the average predicted probability of principal mineralization, averaging the predicted probability of mines, deposits, and occurrences where Cu is present as the main commodity. A higher P_{PM} indicates better performance of the model in predicting Cu occurrences. P_A defines the prospecting area of a region by computing the percentage of area where the average model-predicted probability exceeds 50 %. Upon inspecting the results, we observed that the reduction in the prospecting area is mainly linked to the improved classification of granitic formations as areas not mineralized with Cu—where occurrences are considered improbable. Generally, a lower value indicates better performance in our cases. We highlighted all best values between LA and TL in green, the worst in red, and similar remains black to aid understanding of Tables 5 and 6. The presented values are the means obtained from the 200 realizations, along with their respective standard deviations if available.

Checking Tables 5 and 6, we observe that TL consistently outperforms LA, as evidenced by most metrics labeled in green, with only 9 in red out of 84. Specifically, in Table 5, regardless of the supervised machine learning algorithm used, BS values are significantly lower for TL than LA. This suggests that TL generated more accurate probabilistic predictions than LA for the well-documented area. This superior performance is also reflected in lower RC values, indicating that TL predicted fewer false negatives, meaning a higher proportion of actual positives was correctly identified. As a result, there are higher predicted probabilities associated with mines, deposits, and occurrences where Cu is present as the main commodity. Moreover, standard deviations (values in parentheses) are, for the most part, lower for TL than LA. TL seems to produce a model with less uncertainty than LA, enabling a more precise prediction. Similar conclusions can be drawn for Table 6, which is associated with metrics computed in the remote area.

Note that the gain provided by TL is less pronounced in Table 6 due to only 18 positive data in the validation set, with 9 of them in the same neighborhood (See Fig. 12). This is the reason why some algorithms, especially, XGBoost underperformed when applied using LA. Despite that, on average, the TL approach recognizes that 43.14 % of the area is worth prospecting, while LA mentions 46.29 % in the remote area. This significant reduction of 3.15 % is attributable to better recognition of inert zones. As LA is carried out on a very limited dataset, making this distinction is not easy, which is not the case with TL which has a larger database. Ultimately, the results indicate that using TL can significantly improve forecast accuracy compared to LA, due to its ability to capture more geological signatures of copper mineralization in mineral belts. Importantly, this improvement using TL is not limited to RF but can be extended to several other supervised machine-learning algorithms.

5. Conclusions

This study investigated the impact of TL on mineral prospectivity mapping by training and validating a Random Forest model using five geographically distant mineral belts within the same geological Province. Two criteria were employed to restrict the selection of negative examples, involving the establishment of buffer zones based on distance and spatial association with copper deposits. A total of two hundred potential maps were generated to compare TL with LA. To assess model uncertainty, the mean and standard deviation of the potential maps were calculated, identifying areas where model predictions exhibited the highest stability. Results indicated that the utilization of TL significantly enhanced accuracy and reduced the standard deviation of predictions for mines, deposits, and occurrence locations compared to LA. By augmenting the positive data in the training set, TL demonstrates an improved ability to capture geological signatures of copper mineralization in mineral belts. This was shown in a more accurate prediction of the hotspots in the validation tests, helping for exploration targeting. The case study focused on the Superior Craton geological province, encompassing a significant portion of northeastern Quebec, Canada, in the context of copper prospectivity mapping.

Table 5
Comparison between local approach (LA) and the transfer learning method (TL) based on seven supervised machine learning algorithms for the well-documented area. (RF: random forest; LR: Logistic regression; SVM: Support vector machine; KNN: K-nearest neighbors; AdaBoost: Adaptative boosting; XGBoost: Extreme gradient boosting; ANN: artificial neural networks).

Supervised machine learning method	LA				TL			
	BS (-)	RC (-)	P_{PM} (-)	P_A (-)	BS (-)	RC (-)	P_{PM} (-)	P_A (-)
RF	0.26 (0.03)	0.18	0.73 (0.06)	0.63	0.10 (0.01)	0.17	0.79 (0.06)	0.58
LR	0.51 (0.04)	0.26	0.81 (0.12)	0.57	0.12 (0.05)	0.13	0.85 (0.03)	0.54
KNN	0.32 (0.05)	0.28	0.75 (0.14)	0.47	0.15 (0.02)	0.19	0.82 (0.12)	0.55
SVM	0.45 (0.04)	0.17	0.78 (0.09)	0.59	0.12 (0.05)	0.11	0.87 (0.03)	0.56
ANN	0.52 (0.05)	0.25	0.86 (0.17)	0.61	0.21 (0.06)	0.21	0.92 (0.13)	0.56
AdaBoost	0.38 (0.08)	0.20	0.91 (0.16)	0.61	0.11 (0.02)	0.14	0.96 (0.07)	0.59
XGBoost	0.45 (0.08)	0.26	0.83 (0.33)	0.65	0.11 (0.02)	0.15	0.96 (0.07)	0.59

Table 6
Comparison between local approach (LA) and the transfer learning method (TL) based on seven supervised machine learning algorithms for the remote area. (RF: random forest; LR: Logistic regression; SVM: Support vector machine; KNN: K-nearest neighbors; AdaBoost: Adaptative boosting; XGBoost: Extreme gradient boosting; ANN: artificial neural networks).

Supervised machine learning method	LA				TL			
	BS (-)	RC (-)	P_{PM} (-)	P_A (-)	BS (-)	RC (-)	P_{PM} (-)	P_A (-)
RF	0.08 (0.02)	0.01	0.70 (0.07)	0.40	0.06 (0.02)	0.00	0.75 (0.06)	0.37
LR	0.11 (0.09)	0.09	0.88 (0.13)	0.52	0.06 (0.04)	0.02	0.83 (0.03)	0.51
KNN	0.07 (0.07)	0.04	0.82 (0.10)	0.42	0.06 (0.03)	0.04	0.80 (0.06)	0.41
SVM	0.09 (0.06)	0.02	0.83 (0.06)	0.58	0.06 (0.06)	0.02	0.85 (0.02)	0.51
ANN	0.10 (0.09)	0.09	0.89 (0.13)	0.54	0.08 (0.06)	0.02	0.91 (0.15)	0.51
AdaBoost	0.09 (0.07)	0.03	0.85 (0.25)	0.40	0.06 (0.04)	0.00	0.92 (0.13)	0.36
XGBoost	0.16 (0.01)	0.01	0.57 (0.02)	0.38	0.05 (0.04)	0.00	0.88 (0.08)	0.35

Data Availability Statement:

All python codes can be found on GitHub at: https://github.com/Danlauz/ProspectivityMapping_TL (accessed on December 6, 2023). To download the NTS sheets used in this project, please refer to the SIGEOM website at https://sigeom.mines.gouv.qc.ca/or contact the main author for access to raw and processed data.

CRediT authorship contribution statement

Dany Lauzon: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Erwan Gloaguen:** Conceptualization, Formal analysis, Investigation, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank the Ministry of Natural Resources and Forests, Quebec, Canada (MRNF) for the administrative support at the beginning of this project with the SIGEOM Database. We thank Victor Silva dos Santos for his comments which contributed to the improvement of this manuscript.

Abbreviations

AdaBoost	Adaptative Boosting
ANN	Artificial Neural Network
BS	Brier Score
GSC	Geological Survey of Canada
HF	Hydrothermal Fluids deposits
LA	Local Approach
LR	Logistic Regression
KNN	K-Nearest Neighbors
MERN	Ministry of Energy and Natural Resources, Quebec, Canada
MRNF	Ministry of Natural Resources and Forests, Quebec, Canada
MPM	Mineral Prospectivity Mapping
PA	Prospectable area as the region where the predicted probability exceeds a threshold
PM	Primary Mineralization

(*continued on next column*)

(*continued*)

Abbreviations	
PPM	Average predicted probability of principal mineralization
RC	Recall
RF	Random Forest
SEDEX	Sedimentary Exhalative deposits
SIGEOM	Quebec Government's Spatial Reference Geomining Information System
SM	Secondary Mineralization
TL	Transfer Learning
VMS	Volcanogenic Massive Sulphides
XGBoost	Extreme Gradient Boosting

References

Ali Hosseini, S., Abedi, M., 2015. Data Envelopment Analysis: A knowledge-driven method for mineral prospectivity mapping. Computers & Geosciences 82, 111–119.
Ayer, J. A., Thurston, P. C., and Lafrance, B., 2008, A Special Issue Devoted to Base Metal and Gold Metallogeny at Regional, Camp, and Deposit Scales in the Abitibi Greenstone Belt: Preface: Economic Geology, v. 103, no. 6, p. 1091-1096.
Boadi, B., Sunder Raju, P.V., Wemegah, D.D., 2022. Analysing multi-index overlay and fuzzy logic models for lode-gold prospectivity mapping in the Ahafo gold district – Southwestern Ghana. Ore Geology Reviews 148, 105059.
Breiman, L., 2001. Random Forests: Machine Learning 45 (1), 5–32.
Card, K.D., 1990. A review of the Superior Province of the Canadian Shield, a product of Archean accretion. Precambrian Research 48 (1), 99–156.
Card, K. D., and Poulsen, K. H., 1998, Geology and mineral deposits of the Superior Province of the Canadian Shield.
Carranza, E.J.M., 2009. Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and spatial association with geological features. Ore Geology Reviews 35 (3), 383–400.
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 Geology Reviews 33 (3), 536–558.
Carranza, E.J.M., Laborte, A.G., 2015. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. Ore Geology Reviews 71, 777–787.
Chen, G., Huang, N., Wu, G., Luo, L., Wang, D., Cheng, Q., 2022. Mineral prospectivity mapping based on wavelet neural network and Monte Carlo simulations in the Nanling W-Sn metallogenic province. Ore Geology Reviews 143, 104765.
Cheng, H., Zheng, Y., Wu, S., Lin, Y., Gao, F., Lin, D., Wei, J., Wang, S., Shu, D., Wei, S., Chen, L., 2023. GIS-based mineral prospectivity mapping using machine learning methods: A case study from Zhuonuo ore district, Tibet. Ore Geology Reviews 161, 105627.
Daviran, M., Parsa, M., Maghsoudi, A., Ghezelbash, R., 2022. Quantifying Uncertainties Linked to the Diversity of Mathematical Frameworks in Knowledge-Driven Mineral Prospectivity Mapping. Natural Resources Research 31 (5), 2271–2287.
Esmaeiloghli, S., Tabatabaei, S.H., Carranza, E.J.M., 2021. Spatio-Geologically Informed Fuzzy Classification: An Innovative Method for Recognition of Mineralization-Related Patterns by Integration of Elemental, 3D Spatial, and Geological Information. Natural Resources Research 30 (2), 989–1010.
Ford, A., 2020. Practical Implementation of Random Forest-Based Mineral Potential Mapping for Porphyry Cu–Au Mineralization in the Eastern Lachlan Orogen. NSW, Australia: Natural Resources Research 29 (1), 267–283.
GSC, 2023, Geological Survey of Canada, *in* Natural Resources Canada, G. o. C., ed.
Harris, J.R., Wilkinson, L., Heather, K., Fumerton, S., Bernier, M.A., Ayer, J., Dahn, R., 2001. Application of GIS Processing Techniques for Producing Mineral Prospectivity Maps—A Case Study: Mesothermal Au in the Swayze Greenstone Belt. Ontario, Canada: Natural Resources Research 10 (2), 91–124.

Harris, J.R., Grunsky, E., Behnia, P., Corrigan, D., 2015. Data- and knowledge-driven mineral prospectivity maps for Canada's North. Ore Geology Reviews 71, 788–803.
Harris, J.R., Naghizadeh, M., Behnia, P., Mathieu, L., 2022. Data-driven gold potential maps for the Chibougamau area. Abitibi Greenstone Belt, Canada: Ore Geology Reviews 150, 105176.
Hathway, B., Hudak, G., Hamilton, M.A., 2008. Geologic Setting of Volcanic-Associated Massive Sulfide Deposits in the Kamiskotia Area. Abitibi Subprovince, Canada: Economic Geology 103 (6), 1185–1202.
Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., Azim, M.A., 2022. Transfer learning: a friendly introduction. Journal of Big Data 9 (1), 102.
Jansson, N.F., Allen, R.L., Skogsmo, G., Tavakoli, S., 2022. Principal component analysis and K-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates. Sala Area, Sweden: Journal of Geochemical Exploration 233, 106909.
Lachaud, A., Marcus, A., Vučetić, S., and Mišković, I., 2021, Study of the Influence of Non-Deposit Locations in Data-Driven Mineral Prospectivity Mapping: A Case Study on the Iskut Project in Northwestern British Columbia, Canada, Minerals, Volume 11.
Lachaud, A., Adam, M., and Mišković, I., 2023, Comparative Study of Random Forest and Support Vector Machine Algorithms in Mineral Prospectivity Mapping with Limited Training Data, Minerals, Volume 13.
Lawley, C.J.M., Tschirhart, V., Smith, J.W., Pehrsson, S.J., Schetselaar, E.M., Schaeffer, A.J., Houlé, M.G., Eglington, B.M., 2021. Prospectivity modelling of Canadian magmatic Ni (±Cu ± Co ± PGE) sulphide mineral systems. Ore Geology Reviews 132, 103985.
Lundberg, S.M., Lee, S.-I., 2017. In: A Unified Approach to Interpreting Model Predictions. Curran Associates Inc., Long Beach, California, USA, pp. 4768–4777.
Lusty, P.A.J., Scheib, C., Gunn, A.G., Walker, A.S.D., 2012. Reconnaissance-Scale Prospectivity Analysis for Gold Mineralisation in the Southern Uplands-Down-Longford Terrane. Northern Ireland: Natural Resources Research 21 (3), 359–382.
Mao, X., Wang, J., Deng, H., Liu, Z., Chen, J., Wang, C., Liu, J., 2023. Bayesian Decomposition Modelling: An Interpretable Nonlinear Approach for Mineral Prospectivity Mapping. Mathematical Geosciences 55 (7), 897–942.
Mathieu, L., 2018. Quantifying Hydrothermal Alteration: A Review of Methods. Geosciences 8.
Mathieu, L., 2019. Detecting magmatic-derived fluids using pyrite chemistry: Example of the Chibougamau area, Abitibi Subprovince, Québec. In: Ore Geology Reviews, Vol. 114, p. 103127.
McKay, G., Harris, J.R., 2016. Comparison of the Data-Driven Random Forests Model and a Knowledge-Driven Method for Mineral Prospectivity Mapping: A Case Study for Gold Deposits Around the Huritz Group and Nueltin Suite. Nunavut, Canada: Natural Resources Research 25 (2), 125–143.
Mercier-Langevin, P., Daigneault, R., Goutier, J., Dion, C., Archer, P., 2012. Geology of the Archean Intrusion-Hosted La-Grande-Sud Au-Cu Prospect, La Grande Subprovince, James Bay Region, Québec(1,2). Economic Geology 107 (5), 935–962.
Mohammadpour, M., Bahroudi, A., Abedi, M., 2021. Three dimensional mineral prospectivity modeling by evidential belief functions, a case study from Kahang porphyry Cu deposit. Journal of African Earth Sciences 174, 104098.
MRNF, 2023, Ministère des Ressources naturelles et des Forêts, *in* Québec, G. d., ed.
Nykänen, V., Lahti, I., Niiranen, T., Korhonen, K., 2015. Receiver operating characteristics (ROC) as validation tool for prospectivity models — A magmatic Ni–Cu case study from the Central Lapland Greenstone Belt. Northern Finland: Ore Geology Reviews 71, 853–860.
Parsa, M., Carranza, E.J.M., Ahmadi, B., 2022. Deep GMDH Neural Networks for Predictive Mapping of Mineral Prospectivity in Terrains Hosting Few but Large Mineral Deposits. Natural Resources Research 31 (1), 37–50.
Parsa, M., Maghsoudi, A., 2021. Assessing the effects of mineral systems-derived exploration targeting criteria for random Forests-based predictive mapping of mineral prospectivity in Ahar-Arasbaran area, Iran. Ore Geology Reviews 138, 104399.
Percival, J. A., Skulski, T., Sanborn-Barrie, M., Stott, G.M., Leclair, A.D., Corkery, M.T., and Boily, M., 2012, Geology and tectonic evolution of the Superior Province, Canada, *in* Edited by J.A. Percival, F. A. C., and R.M. Clowes., ed., Chapter 6 In Tectonic Styles in Canada: The LITHOPROBE Perspective, Volume Special Paper 49, Geological Association of Canada, p. 321–378.
Rodriguez-Galiano, V.F., Chica-Olmo, M., Chica-Rivas, M., 2014. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area. Southern Spain: International Journal of Geographical Information Science 28 (7), 1336–1354.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews 71, 804–818.
Ross, P.-S., Boulerice, A., Mercier-Langevin, P., McNicoll, V., 2020. Volcanology, chemo-stratigraphy, geochronology, hydrothermal alteration and VMS potential of the Lemoine Member of the Waconichi Formation. Chibougamau District, Abitibi Greenstone Belt, Québec: Mineralium Deposita 55 (1), 21–46.
Sappin, A.A., Guilmette, C., Goutier, J., Beaudoin, G., 2018. Geochemistry of Mesoarchean felsic to ultramafic volcanic rocks of the Lac Guyer area, La Grande Subprovince (Canada): Evidence for plume-related magmatism in a rift setting. Precambrian Research 316, 83–102.
SIGÉOM, 2022, Geomining Information System, *in* Forêts, M. d. R. n. e. d., ed.
Silva dos Santos, V., Gloaguen, E., Louro, H.A.V., Blouin, M., 2022. Machine Learning Methods for Quantifying Uncertainty in Prospectivity Mapping of Magmatic-Hydrothermal Gold Deposits: A Case Study from Juruena Mineral Province. Northern Mato Grosso, Brazil: Minerals 12 (8), 941.
Soltani, Z., Imamalipour, A., 2022. An improved classification of mineralized zones using particle swarm optimization: A case study from Dagh-Dali ZnPb (±Au) prospect. Northwest Iran: Geochemistry 82 (1), 125850.
Tao, J., Yuan, F., Zhang, N., Chang, J., 2021. Three-Dimensional Prospectivity Modeling of Honghai Volcanogenic Massive Sulfide Cu–Zn Deposit, Eastern Tianshan, Northwestern China Using Weights of Evidence and Fuzzy Logic. Mathematical Geosciences 53 (1), 131–162.
Torppa, J., Nykänen, V., Molnár, F., 2019. Unsupervised clustering and empirical fuzzy memberships for mineral prospectivity modelling. Ore Geology Reviews 107, 58–71.
Williams, N.C., Davidson, G.J., 2004. Possible submarine advanced argillic alteration at the basin lake prospect. Western Tasmania, Australia: Economic Geology 99 (5), 987–1002.
Xi, Y., Li, Y., Liu, J., Wu, S., Lu, N., Liao, G., and Wang, Q., 2023, Application of Analytic Hierarchy Process in Mineral Prospecting Prediction Based on an Integrated Geology-Aerogeophysics-Geochemistry Model, Minerals, Volume 13.
Yin, J., Li, N., 2022. Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping. Ore Geology Reviews 145, 104916.
Zhang, S., Xiao, K., Carranza, E.J.M., Yang, F., 2019. Maximum Entropy and Random Forest Modeling of Mineral Potential: Analysis of Gold Prospectivity in the Hezuo-Meiwu District. West Qinling Orogen, China: Natural Resources Research 28 (3), 645–664.
Zhang, S., Carranza, E.J.M., Xiao, K., Wei, H., Yang, F., Chen, Z., Li, N., Xiang, J., 2022. Mineral Prospectivity Mapping based on Isolation Forest and Random Forest: Implication for the Existence of Spatial Signature of Mineralization in Outliers. Natural Resources Research 31 (4), 1981–1999.
Zhao, J., Sui, Y., Zhang, Z., Zhou, M., 2023. Application of Logistic Regression and Weights of Evidence Methods for Mapping Volcanic-Type Uranium Prospectivity. Minerals 13.
Zuo, R., Carranza, E.J.M., 2011. Support vector machine: A tool for mapping mineral prospectivity. Computers & Geosciences 37 (12), 1967–1975.