

https://dspace.org/

WorkflowInduced Uncertainty in DataDriven Mineral Prospectivity Mapping

S.E Zhang; C.J.M Lawley; J.E Bourdeau; Glen Nwaila; Yousef Ghorbani

https://hdl.handle.net/10539/40831
WOS

Downloaded from a DSpace repository. https://dspace.org/

Natural Resources Research, Vol. 33, No. 3, June 2024 (© 2024)
https://doi.org/10.1007/s11053-024-10322-8

Original Paper

Workflow-Induced Uncertainty in Data-Driven Mineral Prospectivity Mapping

Steven E. Zhang,[1,4] Christopher J. M. Lawley,[1] Julie E. Bourdeau,[1] Glen T. Nwaila,[2] and Yousef Ghorbani[3]

Received 11 October 2023; accepted 26 January 2024
Published online: 2 March 2024

The primary goal of mineral prospectivity mapping (MPM) is to narrow the search for mineral resources by producing spatially selective maps. However, in the data-driven domain, MPM products vary depending on the workflow implemented. Although the data science framework is popular to guide the implementation of data-driven MPM tasks, and is intended to create objective and replicable workflows, this does not necessarily mean that maps derived from data science workflows are optimal in a spatial sense. In this study, we explore interactions between key components of a geodata science-based MPM workflow on the geospatial outcome, within the modeling stage by modulating: (1) feature space dimensionality, (2) the choice of machine learning algorithms, and (3) performance metrics that guide hyperparameter tuning. We specifically relate these variations in the data science workflow to the spatial selectivity of resulting maps using uncertainty propagation. Results demonstrate that typical geodata science-based MPM workflows contain substantial local minima, as it is highly probable for an arbitrary combination of workflow choices to produce highly discriminating models. In addition, variable domain metrics, which are key to guide the iterative implementation of the data science framework, exhibit inconsistent relationships with spatial selectivity. We refer to this class of uncertainty as workflow-induced uncertainty. Consequently, we propose that the canonical concept of scientific consensus from the greater experimental science framework should be adhered to, in order to quantify and mitigate against workflow-induced uncertainty as part of data-driven experimentation. Scientific consensus stipulates that the degree of consensus of experimental outcomes is the determinant in the reliability of findings. Indeed, we demonstrate that consensus through purposeful modulations of components of a data-driven MPM workflow is an effective method to understand and quantify workflow-induced uncertainty on MPM products. In other words, enlarging the search space for workflow design and experimenting with workflow components can result in more meaningful reductions in the physical search space for mineral resources.

KEY WORDS: Mineral prospectivity mapping, Uncertainty, Zn–Pb deposits, Machine learning, Consensus.

[1]Geological Survey of Canada, 601 Booth Street, Ottawa, ON K1A 0E8, Canada.
[2]Wits Mining Institute, University of the Witwatersrand, 1 Jan Smuts Ave., Johannesburg 2000, South Africa.
[3]Joseph Banks Laboratories, School of Chemistry, University of Lincoln, Green Lane, Lincoln, Lincolnshire LN6 7DL, UK.
[4]To whom correspondence should be addressed; e-mail: Steven.Zhang@nrcan-rncan.gc.ca

INTRODUCTION

Mineral prospectivity mapping (MPM) is an activity that derives spatial information regarding the prospectivity of target mineral (or ore) deposits.

1520-7439/24/0600-0995/0 © 2024 Crown

Depending on the approach used, the frameworks that encapsulate the activities in pursuit of MPM are many, and a high-level framework is the exploration information system (EIS) framework (Yousefi et al., 2021). Data-driven MPM is a sub-domain of practice within MPM, whose activities are increasingly formulated into artificial intelligence tasks due to the usefulness of artificial intelligence algorithms, the abundance of data, and the emerging domain of geodata science. The use of artificial intelligence methods in data-driven activities typically adheres to a data science framework, which commonly consists of data collection, data preparation, exploratory data analysis, predictive modeling, and deployment and reporting (Fig. 1) (e.g., Shearer, 2000; Hazzan & Mike, 2023). This type of framework is designed to cater to the characteristics of experimentation, which include: (1) flexibility in methodology (e.g., workflow design), and (2) high-risk of having to execute all stages to reach project feasibility analysis (Fig. 1). There is a hidden type of uncertainty that results from heuristic implementations of the data science workflow, which we call 'workflow-induced uncertainty' (Fig. 1) in the construction of MPM products. In particular, the sensitivity of MPM products is unknown with respect to common workflow design variability, and to what extent workflow uncertainty can be anticipated and mitigated. Fundamentally, this is because workflow design is guided by variable domain performance metrics in the data science framework, but spatial domain characteristics of models are only weakly coupled in a post-hoc manner.

In all sciences, models of reality advance through two modes of scientific inquiry—dissent and consensus. Dissent occurs in the form of new discoveries, breakthroughs and otherwise disruptive information or data that supersedes, invalidates or weakens existing models (Solomon, 1994; e.g., general relativity vs. Newtonian gravity). Consensus reaffirms something that is already known and occurs in the form of replications, validations and successful applications of models, which increase model confidence (Laudan, 1984). Both modes are necessary to achieve scientific advancements, which often occurs through experiments that generate new evidence. Evaluations are then made to assess the degree to which new evidence is dissent or consensus given a previous model. Replicable findings are important to determine the validity of evidence and the degree of either dissent or consensus. Therefore, experimental methodology is designed to maximize outcome replicability. The value of experimental findings is determined through replication, which is a form of scientific consensus (e.g., follow-up studies, replications and post-hoc data analysis). Even dissenting evidence must be replicable to retain scientific value. In the hypothesis-driven domain of science, models being inquired scientifically are causal or physical. In the data-driven domain, models are commonly inferential.

Data-driven MPM is a developing sub-domain (e.g., Zuo et al., 2023), which means that there are no canonical static methodology or standard operating procedures to produce models (which are usually visualized as maps). For example, the construction of evidence layers is seldom comparable across studies (Yousefi et al., 2021; Zuo et al., 2021). The most standardized portion of the data-driven MPM workflow is the data science framework, which recognizes and emphasizes workflow experimentation, guided by model performance. In particular, data science makes the explicit recognition that the applicative branch of this sub-domain is experimental in nature, which is unavoidable for data re-purposing (Grossi et al., 2021). This is entirely the case for MPM, which is the biggest non-ephemeral user of multidisciplinary geodata by its breadth. Experimentation favors method development. However, for method deployment, where a MPM map would be used to guide exploration, rigor in the spatial domain is desirable, such as in the case of mineral resource estimation, which is governed by a rigorous framework on resource assessment (e.g., OSC, 2016; SAMREC, 2016). The typical aspatial data science framework is incomplete and not designed for MPM because most algorithms model data in the variable domain, whose performance is used as guidance in model selection. Spatial connotations of algorithm and model selection cannot be generally addressed within the data science framework because, e.g., hyperparameter tuning of non-spatially aware algorithms cannot be generally and robustly related to spatial characteristics. Applications of data science into each discipline requires addressing challenges that arise from the integration of different bodies of knowledge and traditions (Hazzan & Mike, 2023). Thus far, a geospatial extension has been added to the data science framework for MPM (and similar) tasks, where additional criteria are used after the execution of a data science workflow to contextualize modeling outcomes (Fig. 1). Integration of data science into geosciences has resulted in at least two approaches:

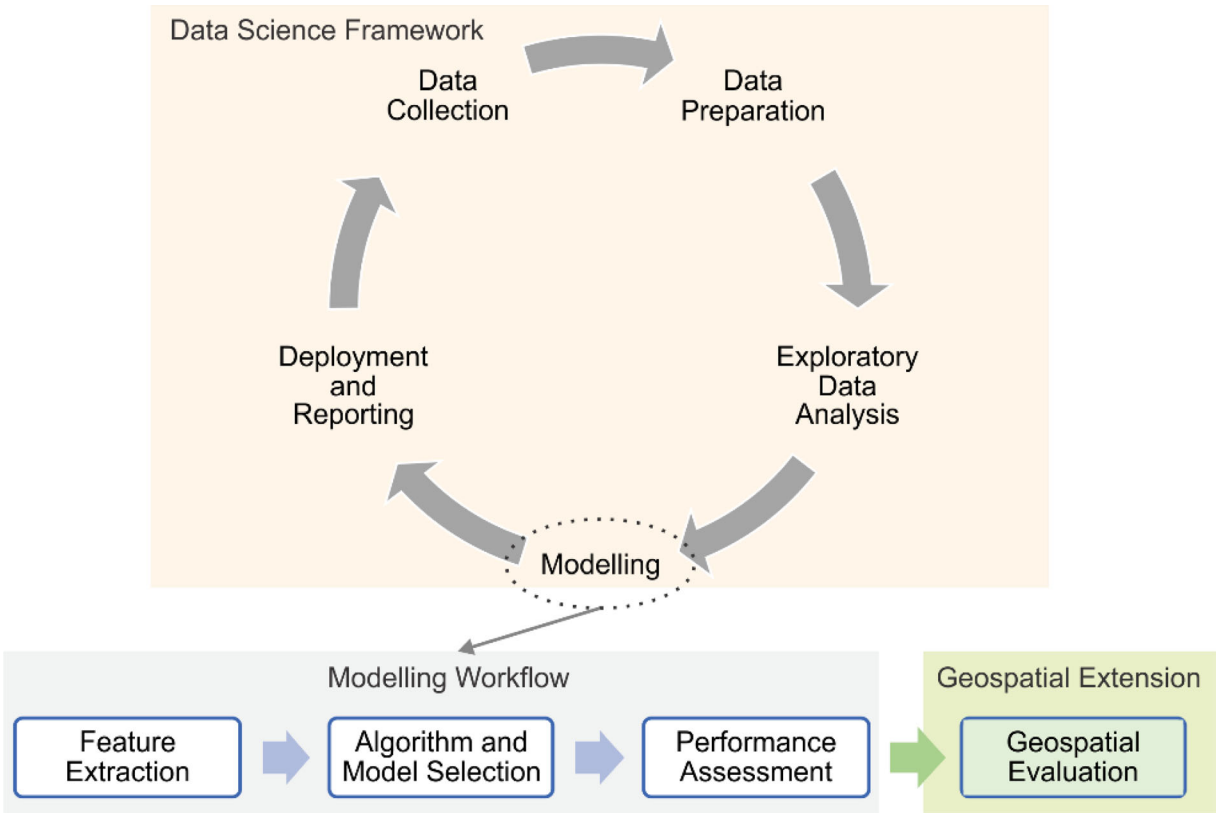


Figure 1. A common data science framework guiding workflow design and execution associated with the generation and usage of predictive models, with a specific geospatial extension to the workflow.

(1) an additional component to data science (e.g., Zuo, 2020), and (2) integration of spatial considerations in data modeling (e.g., Hoffimann et al., 2021). Presently, relationships between variable domain performance (the data science outcome) and spatial characteristics (a geoscience outcome) remain to be fully explored.

Within the modeling stage of the data science framework, several activities must occur whose implementation can be varied, including: (1) details of feature engineering or extraction; (2) algorithm and model selection, including the choice of the performance metric for model selection through hyperparameter tuning; and (3) details of performance assessment (e.g., varying strategies of train-test data splitting) (Fig. 1). The data science framework promotes the variation within any workflow, provided that some benchmark of performance is achieved. From the perspective of data science, there are an unlimited number of combinations of (1)–(3) to achieve a target level of model performance. Since it is impossible to explore all combinations exhaustively, it is unknown whether variations between choices across MPM practitioners result in a substantial difference in prospectivity maps and to what extent are heuristic choices appropriate, in general. This means that it is not possible to achieve scientific consensus of MPM-based exploration models without explicit planning and methodology design. This creates uncertainty that arises from workflow design, which creates unknown impacts on the characteristics of MPM products. The closest known type of uncertainty to workflow-induced uncertainty is judgment-related uncertainty, which was identified for GIS-based approaches, in the context of the human cognitive biases and heuristics (Zuo et al., 2021). Workflow uncertainty is also not directly covered in the aleatoric-and-epistemic decomposition of model uncertainty, although it can impact both (Hüllermeier & Waegeman, 2021).

This study aims to demonstrate that workflow-induced uncertainty exists and is significant, and proposes experimental consensus as a solution. In particular, this study specifically modulates a data-driven MPM workflow in a manner that mimics common variabilities in workflow design, to determine their isolated and systemic effects on the variable and spatial domain characteristics of the resulting models. The modulated components include: (1) feature space dimensionality; (2) the choice of algorithms; and (3) the choice of performance metrics during model selection. We specifically examine the relationship between metrics of model performance during performance assessment in the data science framework and the spatial characteristics of MPM products in the geodata science framework. Our findings are intended to illustrate complex interactions across key tasks that exist in data science workflows for MPM, and their impact on geospatial outcomes.

Our results demonstrate that: (1) it is possible to reach high model performance through many different combinations of feature-algorithm pairings, implying that there are many local minima in workflow design; (2) feature space dimensionality impacts model performance significantly, whose degree is algorithm dependent; (3) the choice of algorithms impact model performance but standard metrics of model performance are insufficient to distinguish the quality of resulting maps as measured by spatial selectivity; and (4) there is an inconsistent relationship between model performance and spatial selectivity, which implies that MPM products should not be solely derived using a data-driven methodology (e.g., the data science framework), but must involve some spatial or knowledge constraints. Based on our analysis, we propose that workflow modulation can be employed to propagate uncertainty to achieve consensus in data-driven MPM.

To ensure replicability and relevance for modern exploration targets, we employed the dataset that was published by Lawley et al. (2021), which contained various data layers and labels pertaining to primarily Mississippi Valley Type (MVT) and clastic-dominated Pb–Zn deposits. It is important to distinguish the purpose of our study from other studies (e.g., Lawley et al., 2022) in that ours is a theoretical/experimental MPM study, whereas, e.g., the Lawley et al. (2022) study is an applicative one. Readers interested in the applicative domain of MPM should refer to Lawley et al. (2022) for applied MPM products.

REVIEW OF DATA-DRIVEN MPM WORKFLOWS

MPM can be conducted in a manner guided by knowledge, data or some combination thereof (Yousefi et al., 2021). In geodata science-based MPM, machine learning algorithms are commonly used to model relationships between covariates and target labels because MPM is a particularly data-rich domain within the geosciences and machine learning as a discipline is primarily intended for big data analysis (Zuo, 2020; Yousefi et al., 2021). However, the extension of data science into geodata science is more recent and ongoing (e.g., Zuo, 2020; Yousefi et al., 2021) (Fig. 1). Within MPM, and for strictly the purpose of data analysis, machine learning is used to primarily: (1) identify mineralization-related anomalies through unsupervised learning (e.g., Nwaila et al., 2022; Zhang et al., 2022b); (2) predict targets that are similar to known occurrences through supervised learning (e.g., Zuo & Carranza, 2011; Zhang et al., 2021; Senanayake et al., 2023); and (3) predict targets using reinforcement learning (e.g., Shi et al., 2023). Outside of data analysis, machine learning is also beginning to be used in: (1) data generation (e.g., Zhang et al., 2022b; Bourdeau et al., 2023); (2) data processing (e.g., Song et al., 2020; Nwaila et al., 2023; Zhang et al., 2023); and (3) simulations (e.g., Song et al., 2021). Shallow learning and statistical approaches are very common (Zuo, 2020). Deep learning is common in MPM because of its lack of a-priori considerations on the statistical properties of data and the availability of sizable datasets at regional to national scales (e.g., Xiong et al., 2018; Li et al., 2020; Sun et al., 2020; Shi et al., 2023). Within the deep learning algorithms, autoencoders (e.g., Chen, 2015; Xiong et al., 2018) and convolutional neural networks (Li et al., 2020; Sun et al., 2020) are prominent (Zuo, 2020). For autoencoders, this is because they are the most general-purpose dimensionality reduction and data reconstruction tools, which means they could be used to both extract features and detect high-dimensional anomalies. For convolutional neural networks, this is because they are the most general-purpose learning algorithms that are applicable to rasterized (image-like) data, can perform automated feature extraction, and can extract spatial relationships in addition to those in the variable domain.

This study solely focuses on classification-based data-driven MPM using supervised machine learning algorithms in the variable domain (or spectrum-

based methods, see Zuo & Xu, 2024). This class of approaches use training data with labels in the form of known mineral occurrences (e.g., deposits, prospects and showings) that are used to guide model construction (Yousefi et al., 2021). Supervised data-driven methods are commonly biased by the locations of training data labels (Yousefi et al., 2021). In addition, the high complexity of geological systems combined with an unknown lower-bound number of labels renders it easy to produce models that are overfitted, particularly for deep learning algorithms (Dietterich, 1995; Porwal et al., 2004; Coolbaugh et al., 2007; Skabar, 2007; Srivastava et al., 2014; Chen, 2015; Porwal et al., 2015; Zuo et al., 2019, 2021; Yousefi et al., 2021). Consequently, models that are produced using data-driven methods carry exploration biases and uncertainties (Yousefi & Nykänen, 2016). The formulation of MPM into artificial intelligence tasks is an emerging domain of research because there are generally unsolved challenges including: (1) the selection of algorithm, architecture and optimal parameters; (2) training data sufficiency, achieving class balance and representativity of negative labels; and (3) rigorous utilization of data science (Yousefi et al., 2021). In addition, there are additional unsolved issues, e.g., of the spatial transferability of models and performance expectations; specifically, the lack of geostatistical learning algorithms (e.g., Hoffimann et al., 2021).

An absolute scientific requirement is that measurements or findings must be associated with a notion of uncertainty. Here, we summarize uncertainty in MPM products in strictly the data-driven domain. Data-driven MPM products exhibit at least two classes of uncertainty: (1) aleatoric, or data related; and (2) epistemic, or model related (Zuo et al., 2021). Aleatoric uncertainty arises from imperfect data, which can be further categorized into those related to the spatial or variable domains. It is important to distinguish between ''data quality'' as it is referred to in MPM literature (e.g., Zuo et al., 2021) and in data management (e.g., DAMA DMBOK; Henderson et al., 2017). The former aligns closer with GIS terminology, while the latter is a broader framework that aligns with data management and analytics. The latter definition is constructed using the delineation of roles along the data lifecycle, by treating data as a product (e.g., similar to a car from design to manufacturing processes) that is passed from data generators to users. In this definition, characteristics of data from generation to management that adversely impacts analytics performance are considered quality-related. We hereon standardized the terminology as per DAMA's definition (Henderson et al., 2017) because this study considers data-driven MPM using machine learning algorithms as a specific use-case of artificial intelligence, which is a form of advanced analytics. Therefore, completeness, resolution and availability are part of data quality, which is different than the GIS definition (Zuo et al., 2021). In this definition, spatial domain uncertainty can be affected by (non-exhaustively): coverage and coverage rate, especially relative to the areal coverage of litho-diversity classes, and imprecision of coordinate measurements. Variable domain uncertainty can be affected by (non-exhaustively): metrological characteristics, such as measurement accuracy and precision; sampling practices; and geostatistical characteristics, such as the nugget and random noise effects. Epistemic uncertainty relates to how models are typically undercomplete relative to the full behavior of mineral systems, due to model capability, availability of algorithms, knowledge of target behavior and naturally irreducible variability (e.g., complex system behavior). For example, algorithms that result in models with greater degrees of freedom, such as neural networks, are capable of modeling more complex phenomena than simpler algorithms, such as linear regression.

In addition to the preceding classes of uncertainty, there is an additional class (3) that captures judgment-related uncertainty, which is the result of cognitive heuristics, experiential bias and other primarily user-related effects, in a GIS context (Zuo et al., 2021). This class has no direct equivalent in a pure data science framework (Fig. 1) because this framework was purposefully created to remove user judgment by selecting models based on objective metric scores and permitting workflow modulations to achieve higher scores. Metric-driven search for model candidates creates a collection of models, each containing error that is decomposable into bias, variance and irreducible error (Kohavi & Wolpert, 1996). However, judgment-related uncertainty exists in the geodata science framework (Fig. 1) because the spatial basis for model selection is EIS or GIS based (e.g., Yousefi et al., 2021), but model construction through hyperparameter tuning is not. Therefore, performance as assessed in the variable domain using common scoring metrics is decoupled from spatial characteristics or performance. Consequently, it is possible to produce equiprobable

models, whose model performance in the data science framework are comparable, but whose spatial domain characteristics are not. This is compounded by the fact that only a tiny minority of models per workflow are assessed in the geospatial sense (e.g., the vast majority of models explored through grid search are not assessed in the spatial domain). Therefore, variations in workflow design creates uncertainty in MPM products because user judgment or perceived best practices (e.g., interpolation) guide the implementation of at least some stages of the geodata science framework (Fig. 1). In addition, workflow tuning based on metric scores implies that models are produced using a workflow-wide gradient descent to the best possible combination of choices. However, given the large exploration space for workflow design, reaching the global minimum is impossible in general, and instead, equiprobable models pertain to local minima in workflow design. This type of uncertainty would not be an issue, if spatial domain characteristic is monotonically dependent on variable domain performance, which means optimization in one domain is equivalent to the optimization in the other. However, this has never been examined empirically nor is it theoretically obvious. Presently, it is unknown how typical variability in the construction of data science-based MPM workflows translate into spatial characteristics.

Within the domain of supervised artificial intelligence-based MPM, we consider algorithm selection as the biggest documented source of variability. This is because performance evaluations are typically made on an experimentation of algorithm selection, while controlling other portions of the workflow. This is exemplified by a sizable body of literature, whose raison d'être is to demonstrate the effectiveness of novel algorithms in MPM tasks (Chen & Wu, 2017; Xiong et al., 2018; Chen et al., 2020; Wang et al., 2020; Yang et al., 2022; Yin & Li, 2022; Zuo et al., 2022; Gharehchopogh et al., 2023; Li et al., 2023; Yin et al., 2023). There also have been recent efforts to examine specifically this type of uncertainty, in a GIS knowledge-driven framework (Daviran et al., 2022). The range of all possible algorithms is unknowable because there are emerging algorithms and variations of existing ones, either as architectural modifications (e.g., changes in neural network architecture) or as add-ons (e.g., optimization algorithms; Chen et al., 2020; Yin & Li, 2022; Gharehchopogh et al., 2023). An empirical analysis revealed that algorithms used by various authors include (non-exhaustively): Bayes network (Porwal et al., 2006; Yin & Li, 2022); logistic regression (Agterberg & Bonham-Carter, 1999; Carranza & Hale, 2001; Karbalaei Ramezanali et al., 2020; Lin et al., 2020; Zhang et al., 2022c); support vector machines (Zuo & Carranza, 2011; Zhang et al., 2021; Senanayake et al., 2023); tree-based methods, such as random forest, extra trees and XGBoost (Chen & Wu, 2019; Sun et al., 2019; Zhang et al., 2022a); artificial neural networks, such as extreme learning machines (Chen & Wu, 2017); deep learning methods (Xiong et al., 2018; Wang et al., 2020; Yang et al., 2022; Zuo et al., 2022, 2023; Li et al., 2023; Yin et al., 2023; Zuo & Xu, 2023); and reinforcement learning (Shi et al., 2023). There are also applicative MPM studies that employed ensemble learning, which is an approach to improve outcome reliability by integrating the output of multiple independent models (e.g., Senanayake et al., 2023; Shetty et al., 2023). The diversity of algorithms reflects a diversity of practitioners, scale of MPM tasks, computational and data capabilities, and algorithmic complexity choices, all of which emphasize the experimental nature of data-driven MPM. Algorithm choice primarily controls the explanatory power of the model (but also model complexity and generalizability), which implies that modulating algorithmic choice impacts epistemic uncertainty (Hüllermeier & Waegeman, 2021).

A second major source of variability in data-driven workflows is the construction of covariates, which are referred to as 'evidence layers' in the geospatial realm. Covariate data can range from mono-disciplinary, such as geochemical (e.g., Zhang et al., 2022b) or spectral (e.g., Nwaila et al., 2022), up to sizable combinations of qualitative and quantitative geoscientific data (e.g., Lawley et al., 2022). The variability encountered in covariate construction falls into two major categories: (1) the number of covariates; and (2) their quality. Category (1) variability primarily controls the threshold between overfitting and underfitting that differs on a per-algorithm basis (through the curse of dimensionality; Jia et al., 2022; Márquez, 2022), which further implies that the selection of data and feature engineering impacts algorithm selection. The presence of category (2) variability is theoretically easy to deduce because factors such as the support type (in a geostatistical sense), the quality of interpolation, the data coverage and coverage rate (e.g., density of samples and their distribution over space), as well as metrological properties (e.g., measurement accuracy and precision) all impact the quality of models. The

impact of data quality on models is a part of aleatoric uncertainty (Hüllermeier & Waegeman, 2021). However, the impact of data quality on downstream MPM products is not fully understood because such interactions are complicated to model in general and even for specific applications, requires repetitive executions of the data science framework, through uncertainty propagation (Fig. 1) (e.g., Yang et al., 2023). Consequently, pre-data modeling processes typically occur in a feed-forward fashion. For example, spatial interpolation of sparse data into evidence layers are often performed using heuristic or best practices, although the rationale is inconsistently documented (e.g., Senanayake et al., 2023). Creation of evidence layers in this manner is considered feed-forward because much of the best practices predate the coupling of traditional geoscientific data to artificial intelligence algorithms. Consequently, it is not generally known whether such best practices are still optimal in a data re-purposing context because data usage methods have changed. Purely data-related workflow variability is outside of the scope of this study because this study uses a published and rasterized dataset. However and for example, varying how data are collected and processed clearly contributes to aleatoric uncertainty and could theoretically be varied to modulate their effects on MPM products (e.g., partially via simulations, see Yang et al., 2023). A pre-data modeling task that is often performed is dimensionality reduction because of the high dimensionality of evidence layers relative to the paucity of training labels (in supervised approaches). If left unattended, using a native number of evidence layers can lead to overfitting because the degrees of freedom of complex models are not fixed using a statistically robust number of samples. Feature space dimensionality reduction occurs using algorithms that range from knowledge-based feature engineering, simple principal component analysis to knowledge-constrained variational autoencoders (e.g., Zuo et al., 2022; Senanayake et al., 2023). Technically, the modulation of feature space dimensionality impacts both variability categories (1) and (2) because this process affects both the number of features available for predictive modeling and the quality of the features through a controlled loss in feature-explanatory power.

Popular choices of performance metrics include those that measure model discrimination power, such as the area-under-the-curve of the receiver operating function (AUC-ROC), and those that measure the quality of predictions, such as the accuracy and F1 metrics. Metrics are intended to be a robust criterion to guide model selection or hyperparameter tuning. However, in the geodata science extension (Fig. 1), this intention is weakened because spatial domain metrics, such as the prediction-area curve (Yousefi & Carranza, 2015a, 2015b) does not directly guide hyperparameter optimization (or model selection; e.g., Yin & Li, 2022). Part of the issue results from the fact that training datasets are not necessarily (and generally are not) spatially contiguous and of sufficient extent, such that spatial characteristics could be robustly determined during hyperparameter optimization. Consequently, spatial metrics are typically computed on the testing dataset (or more appropriately, the whole dataset), which means that spatial performance assessed post-hoc of hyperparameter optimization cannot be used to rigorously guide model selection but only model performance assessment in the spatial domain (e.g., Yin & Li, 2022). Therefore, it is not clear the relationships between the choice of variable domain metrics and spatial characteristics. The choice of feature space dimensionality is also similarly weak because the extent to which the explanatory power of covariates is preserved during feature engineering occurs along a smooth continuum. Once this choice is made, it is not generally documented to have been revisited explicitly, which implies that the impact of this choice on downstream metrics is not generally intelligible. Consequently, the complex interactions between feature space dimensionality and algorithm performance, particularly the relationships between variable (in a pure data science framework) and spatial domain performance (in a geodata science framework), remain to be explored.

METHODOLOGY

Data Sources

The dataset employed in this study was published and made digitally available by Lawley et al. (2021) as a single file or as a compilation of GIS-compatible files from McCafferty et al. (2023) (Fig. 2). The dataset was managed to facilitate its reuse and re-purpose by data scientists because the master data, metadata and reference data files are fully machine readable and full documentation exists to guide non-geoscientists to understand the data. By employing this dataset, we were able to experimentally control the data collection to preparation stages of the data science workflow.

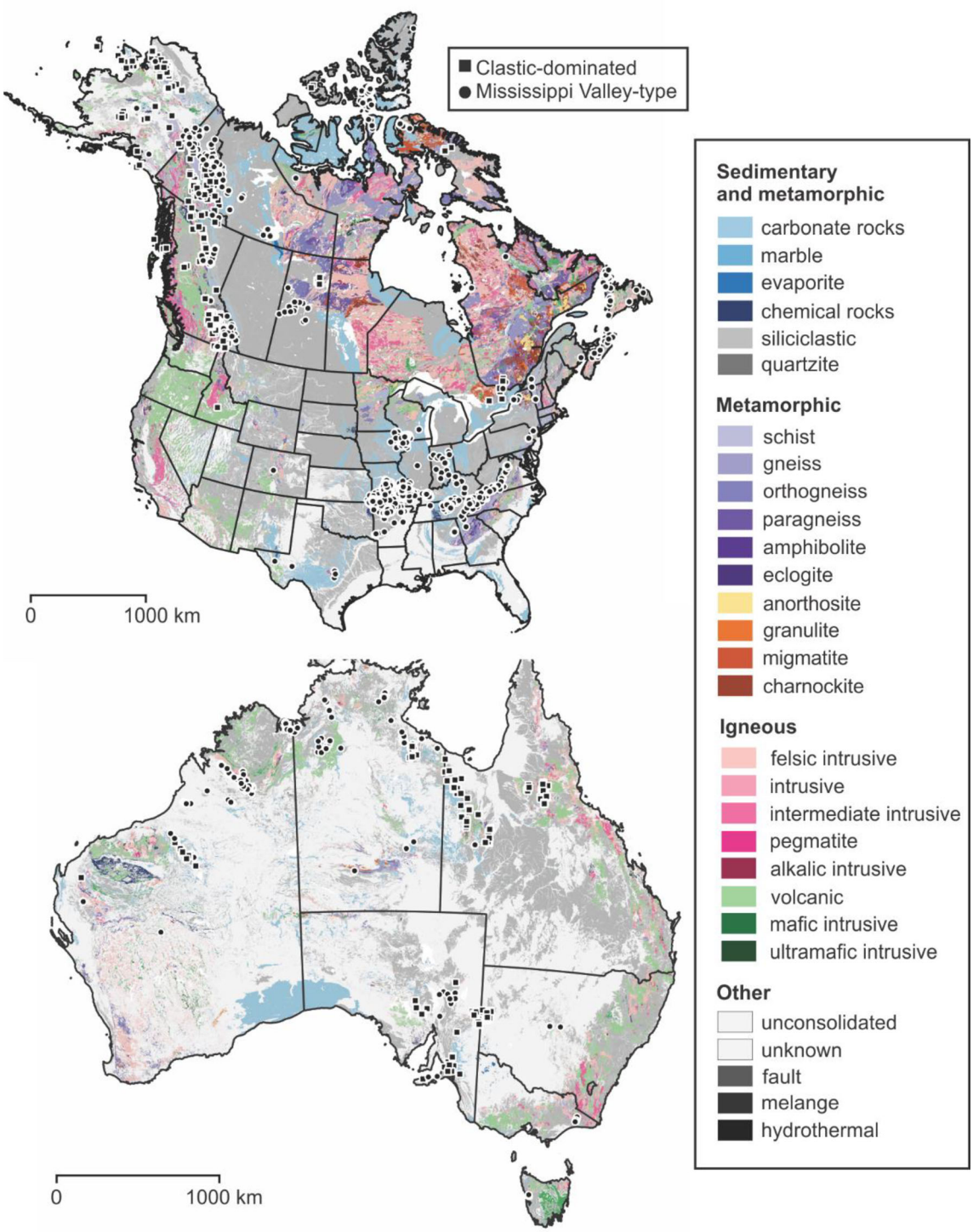


Figure 2. Geological maps of Canada and the United States (top), and Australia (bottom). Figure was modified from Lawley et al. (2021). The H3 cells are color-coded using the hierarchical rock classification developed as part of Lawley et al. (2021), with data sourced from multiple national, state, provincial/territorial databases. Rock subtypes were used during mineral prospectivity mapping as mappable proxies for the sources and traps of mineral systems. Known Pb–Zn Mississippi Valley-type and clastic-dominated deposits and mineral occurrences used for training are shown for reference.

Here, we provide a description of the dataset (Fig. 2). Bedrock geological data for the United States were sourced from the State Geologic Map Compilation (SGMC) digital database covering the conterminous United States (Horton et al., 2017), as well as the Geologic Map of Alaska (Wilson et al., 2015). In the case of Canada, the geological datasets are a compilation of 20 previously published national, provincial and territorial geological map databases (Lawley et al., 2021). It is important to note that geological data from the United States and Canada should be considered a collection of individual bedrock geology maps with highly variable mapping scales (1:50,000 to 1:5,000,000) and boundary artifacts (e.g., induced by mapping subjectivity). In contrast, geological data for Australia are seamless and was extracted from the 1:1 million scale national bedrock geology database (Raymond et al., 2012). Lithological information from all 23 source maps was re-categorized into four main types (sedimentary, igneous, metamorphic, and other) and 31 subtypes for the purposes of prospectivity modeling (Fig. 2). New data dictionaries were also used to identify the presence or absence of up to 17 geological properties (such as coarse clastic, fine clastic, calcareous, carbonaceous, evaporitic, cherty, red beds, sedimentary, ultramafic to mafic composition, intermediate composition, felsic composition, pegmatitic, alkalic, igneous, schistose, gneissose, and anatectic) from the available geoscientific text data.

Geochronological data were reformatted and compiled from the 23 geological map compilations described above. They were then combined with plate tectonic models to estimate the paleo-latitude and -longitude of rocks at the time of their deposition or emplacement (Scotese, 2021). The fault compilation was based on global sources (Chorlton, 2007; Styron & Pagani, 2020), national databases (Raymond et al., 2012), and the 1:5,000,000 scale Geologic Map of North America digital database (Reed et al., 2005) (Fig. 2). Duplicate faults from these different data sources were not removed prior to converting fault traces into a proximity surface. Proximity calculations were also performed for passive margins and the point locations of carbonaceous sedimentary rocks (e.g., black shales) using the global compilations from Bradley (2008) and Granitto et al. (2017), respectively.

Geophysical datasets represent the second major source of information for prospectivity modeling. Seismic datasets were sourced from a range of survey types, including active controlled-source seismic refraction and passive teleseismic surveys. Depth estimates for the seismogenic Moho were extracted from national datasets specific to each region, i.e., Canada (Schetselaar & Snyder, 2017), the contiguous United States (Shen & Ritzwoller, 2016), Alaska (Zhang et al., 2019), and Australia (Kennett et al., 2011). Global models, such as Szwillus et al. (2019), Reguzzoni & Sampietro (2015), and Laske et al. (2013) were also used as secondary sources for Moho depth estimates. Seismic velocity data for the upper mantle and estimates for the lithosphere–asthenosphere boundary were sourced from studies conducted by Debayle et al. (2020), Hoggard et al. (2020), and Priestley et al. (2018). The spatial resolution of the source seismic datasets is highly variable (e.g., seismic stations may be separated by 10–100s of km), but the depth estimates for the seismogenic Moho and lithospheric–asthenosphere boundary were expected to be within a few kilometers of their true value (Snyder et al., 2018).

Gravity datasets were sourced from satellite, airborne, and ground-based surveys. Satellite gravity data is based on the European Space Agency (ESA) Gravity Field Steady-State Ocean Circulation Explorer (GOCE) mission and its derivative products, which focus on highlighting the curvature and shape of gravity anomalies, as described in Ebbing et al. (2018). The Bouguer anomaly grid is based on national gravity databases for Canada, Alaska, the contiguous United States, and Australia (Phillips et al., 1993; Saltus et al., 2008; Geological Survey of Canada, 2017; Lane et al., 2020). Individual stations within this gravity measurement network tend to be distributed every few kilometers but can be more widely spaced for the more remote parts of North America and Australia. Multiple derivative products were then calculated to highlight density variations at different crustal depths, including the horizontal gradient magnitude (HGM) and an upward continued version of the Bouguer anomaly grid. In addition to these grids and their derivatives, we identified points tracking the maxima of the HGM gravity values for both the Bouguer gravity and its upward-continued field following the approach outlined by Phillips (2007). The proximity to these tracks, often referred to as gravity ''worms'', were used as proxies for identifying steeply dipping geological structures and the contacts between rocks with varying densities.

Magnetic datasets were combined from both modern sources (Alaska Division of Geological and Geophysical Surveys, 2016; Miles & Oneschuk, 2016) and legacy surveys, such as the North America

magnetic map (Ravat et al., 2009), to create an updated and merged residual magnetic anomaly map for North America. The spatial resolution of these source geophysical surveys is highly variable, and, in the case of aeromagnetic data, can include line spacings ranging from hundreds of meters to several kilometers. A differential RTP (reduction-to-pole) methodology included with the Geosoft Montaj software was used to make the adjustment for the magnetic north pole, as described by Arkani-Hamed (2007). An identical workflow was followed to produce an RTP grid and its derivative products from the residual magnetic anomaly map of Australia. In addition, several derivative products were calculated from the RTP grids, following the procedures outlined in Phillips (2007). These calculations included the first vertical derivative (1VD) to assess and map magnetic variations in the shallow crust. Shallow magnetic sources were further highlighted by calculating the horizontal gradient magnitude (HGM) of the pseudo-gravity field and the proximity to the paths of HGM maxima, often referred to as magnetic ''worms''. Deep magnetic sources are based on long-wavelength anomalies from the RTP grids using a matched filtering technique described by Syberg (1972) and implemented with code developed by Phillips (2007). These long-wavelength RTP anomalies were then transformed into a pseudo-gravity field, and the HGM was calculated from this field to accentuate the boundaries and edges of deep magnetic sources. The proximity to long-wavelength ''worms,'' which trace the HGM maxima, served as proxies for mapping the edges of deep magnetic sources within the deep-to-mid-crust.

The dataset was published in the H3 Discrete Global Grid System (DGGS), an open-source software developed by Uber Technologies Inc. (2020) to convert latitude and longitude coordinates of the source geological and geophysical datasets into unique H3 addresses (Fig. 2). The H3 DGGS is a hierarchical and hexagonal grid system, which offers global coverage at multiple resolutions. For spatial indexing, this study used resolution 7 of the H3 DGGS, resulting in 98,825,162 unique H3 addresses worldwide. At this resolution, the average hexagon covers an area of 5.16 km^2, with average edge length of 1.22 km.

Data Processing

Data pre-processing is trivial because of the high quality of the dataset. There were a total of 82 evidence layers, of which 31 contained null values for at least one data record. These layers were removed because this study does not aim to examine the variability introduced by imputation methods, which is a part of data preparation. Additionally, the intention of this study was not to provide updated prospectivity maps, which was completed by Lawley et al. (2022) for the same exact dataset. Consequently, the employed data pre-processing may result in a loss of model performance or spatial realism of derived maps, under the condition that: (1) data imputation is feasible and benefits predictive modeling; and (2) the affected data layers are useful additional covariates of the target label. After the removal of partially empty evidence layers, the data were re-scaled to span an equal range. The labels in the dataset describe two types of Zn–Pb deposits, that is, MVT and clastic-dominated (CD) deposits. For the purpose of our study, we wished to maximize the amount of training data possible. Therefore, it was irrelevant to the study's purpose, whether the targets were MVT or CD (or any other deposit type). Consequently, we merged the two deposit types into a single class, implying that MPM is targeting all Zn–Pb deposit types. Training data were purposely generated to achieve class balance. Hence, there was a total of 2474 positive labels and the same number of negative labels, which were randomly sampled from the much larger proportion of presumed negative labels. However, negative training labels were kept consistent across workflow variations to control training data variability. This was sufficient for the purpose of pure (as opposed to applied) MPM research to understand workflow variability that excludes variability induced by data label generation. In addition, with the exception of class balancing, this was identical to the original methodology used in Lawley et al. (2022) because true negatives (areas devoid of mineralization) are generally unavailable across all areas, due to a limited availability of borehole coverage.

Data Science Workflow and Algorithm Descriptions

The workflow contained three key stages related to the experimental variables: (1) feature extraction and feature space dimensionality-modulation using autoencoders; (2) predictive modeling using machine learning algorithms; and (3) statistical analysis and geospatial visualization. For (1), there is a collection of algorithms in the data-driven domain

that are theoretically useful (Jia et al., 2022), but research of their empirical effects on MPM is exceedingly rare (Wang et al., 2022). For our purpose, the experimental constraint to maximize the amount of information contained in the covariates for every feature space dimensionality implies that we require a general-purpose feature extraction and engineering algorithm that is agnostic to downstream algorithms. Therefore, we cannot rely on feature elimination methods because their outcomes are optimized to downstream algorithms. Other popular methods in MPM include mainly principal component analysis (PCA) (e.g., Wang et al., 2022). However, PCA is not the most general because it assumes linearity of relationships in data, which, if violated, can lead to counter-productive results (Abbott, 2014). Consequently, for dimensionality reduction, we employed autoencoders, which are generalizations of PCA, because among other characteristics, they make no assumptions on the types of relationships exhibited by data (Kramer, 1991). Therefore, we ensured that features extracted from the evidence layers were the most compact, decorrelated (linearly and nonlinearly), information-dense and useful for all downstream algorithms. This enabled us to ensure that feature space dimensionality can be explicitly controlled and that information is maximally retained for any feature space dimensionality. Consequently, we were able to modulate feature space density by changing the number of extracted features to study its impact on algorithms and maps.

Autoencoders are a general-purpose dimensionality reduction tool based on artificial neural networks (ANNs) in the context of unsupervised feature extraction via deep learning. The classic function of autoencoders is to encode data into a compact latent space representation, such that reconstructed data and input data are as similar as possible (Kramer, 1991). ANNs mimic biological neural networks and feature multiple layers of varying numbers of neural nodes, which in the case of feed-forward ANNs, are connected in a single direction from the input side of the network to the output. Abstract information flows along the connections. At each node, an activation function computes the nodal output given the sum of all weighted inputs from other nodes. Typical activation functions include the rectified linear (relu), hyperbolic tangent (tanh), logistic and linear functions. Where autoencoders are used for dimensionality reduction, they are a generalization (e.g., nonlinear and non-parametric) of PCA (Kramer, 1991). The bottleneck of the autoencoder, which is the narrowest layer of the ANN, also known as the coding layer, is purposely designed to be narrower than the dimensionality of the input data, such that the typical autoencoder is undercomplete—meaning that the neural network is unable to fully replicate all information contained in data. Hence, by modulating the coding layer size, autoencoders can discard variable amounts of information in the data. This behavior is similar to truncating the number of principal components. Reconstructed data exhibit an increasing loss of noise, data and physical anomalies and increasingly more common events, with decreasing coding layer size. Consequently, known uses of autoencoders include image denoising, geochemical anomaly detection and seismic data denoising (e.g., Gondara, 2016; Luo et al., 2020; Saad & Chen, 2020, respectively). Key hyperparameters of ANNs include the number of hidden layers and their size, the choice of the activation function, the ability to stop training early if performance fails to improve (early stopping), regularization strength, and the random number generator seed to control the initialization of all connection weights.

Shallow learning algorithms are suitable for predictive modeling but are incapable of automated feature extraction. Unlike deep learning, their usage necessitates discipline-specific expertise and knowledge to craft effective features for the task and algorithms considered. In this study, we explored a variety of common shallow learning algorithms with a range of algorithmic complexity that include: logistic regression (LR; Cramer, 2002); *k*-nearest neighbors (kNN; Tikhonov, 1943; Fix & Hodges, 1951; Cover & Hart, 1967); Gaussian process (GP; Rasmussen & Williams, 2006; Kotsiantis et al., 2007); artificial neural networks (when not used for deep learning, ANN; Curry, 1944; Rosenblatt, 1961; Rumelhart et al., 1985; Hastie et al., 2009; Lemaréchal, 2012); support vector machines (SVM); random forest (RF) and adaptive boosting of decision trees (AdaBoost or AB; Ho, 1995; Breiman, 1996a, 1996b; Freund & Schapire, 1997; Breiman, 2001; Kotsiantis, 2014; Sagi & Rokach, 2018); and extra trees (ET; Geurts et al., 2006). It is worth mentioning that ANN can be used for both shallow and deep learning purposes and, where it is used for both feature extraction and predictive modeling, the combination is considered deep learning.

In this study, we forwent an exhaustive technical description of the machine learning algorithms for brevity but provide a general description of them

and their key hyperparameters. The LR algorithm is a parametric method that attempts to fit data labels to a logistic function. In the classic case, the output of the model is binary, which suits classification tasks. Its key hyperparameters include the type of regularization, which could be none, L1-norm, L2-norm or a combination (elastic net) thereof. The strength of regularization is controlled by a parameter C. The kNN algorithm is a non-parametric method that predicts the label of a sample using an average of similar samples, which is computed through a distance metric (typically the L2-norm). Its key hyperparameters include the number of neighbors (k) and distance weighting, which could be a variety that include uniform or inversely weighted by distance. GP is a non-parametric feature interpolation-based method that probabilistically assigns labels to samples by attempting to model the underlying stochastic process using a kernel. The kernel specifies the covariance function of the data. The type of kernel and its parameters are key hyperparameters. Common kernel choices include the radial basis function (RBF), which has two typical parameters that specify the length scale and amplitude. The SVM algorithm is a non-probabilistic linear method that attempts to maximize the separation distance between classes. Because it can be formulated using inner products, it is often used with kernels to effectively expand the dimensionality of the problem without data transformation and seek a decision boundary in higher dimensional space, which when projected to the original dimensions of the problem, becomes a nonlinear decision boundary. Key SVM hyperparameters include: the choice of the kernel, which could be RBF, linear or polynomial; C, which defines a penalty for misclassifying support vectors and higher values increase the decision boundary complexity; and γ, which specifies the nonlinear kernel's coefficients.

Ensemble methods include RF, AB and ET. Although technically, any weak classifier could be ensembled to produce a strong classifier, RF, AB and ET typically refer to ensembles of decision trees, which are flowchart-like hierarchical structures that partition the trees recursively. Learning occurs to partition the trees through splitting branches (at nodes). The splitting is metric-driven to maximize the difference between the resulting leaves. RF utilizes bagging to create a forest of de-correlated trees, from which the averaged output is the prediction. De-correlation occurs through bootstrap sampling of features. The maximum number of features per tree, the splitting metric, the number of trees, and the minimum number of samples per split are model hyperparameters in addition to the tree depth parameter that is inherited from decision trees. AB uses adaptive boosting (Freund & Schapire, 1995), which is a type of predictor–corrector algorithm in the sense that trees are constructed through adaptation, which adjusts the weights of subsequent trees to focus on cases that are more difficult. The output in AB is a weighted sum of individual trees. In addition to hyperparameters for RF, the rate of adaptation is an AB-specific hyperparameter. ET is similar to RF but randomizes thresholds instead of optimizing them for splitting, and retaining the best thresholds as the splitting rule. The hyperparameter grid for all algorithms is provided in Table 1. The hyperparameter ranges are not intended to be exhaustive but are intended to cover the probable usage of each algorithm in this task. In addition, the hyperparameter ranges were iteratively refined by observing model choices (e.g., if a model is consistently at the end of a parameter range, that parameter range would be extended, if possible, to increase model diversity).

Model selection occurs through cross-validation, whose strategy is an experimental control. Because the data were abundant, we employed four-fold cross-validation. We used a variety of metrics to assess model performance, which included accuracy, class-weighted F1 score and AUC-ROC (Fawcett, 2006). The F1 score is a harmonic mean of precision and recall. The ROC (receiver operating function) is a curve that describes the relationship between true positive rate and false positive rate at a range of discrimination thresholds, whose integral is the AUC (area under the curve)-ROC score. In essence, the AUC-ROC measures the degree of cross-over of the predictions given a model. However, the AUC-ROC metric can give misleading results in class-imbalanced cases. Consequently, the AUC-ROC metric is better for examining the quality of a model, whereas the class-weighted F1 score is robust to class imbalance and is better for assessing the quality of predictions (e.g., outputs of a deployed model). AUC-ROC scores are easy to interpret; e.g., a score of 0.5 implies that the model is no better than a random guess. The AUC-ROC score is desirable for classification tasks because it is classification-threshold-invariant and describes the model's discrimination quality. This further implies that MPM models derived from different implementations of the data science framework can be compared purely

Table 1. Hyperparameter grid for all algorithms

Algorithm	Parameter range
ANN (autoencoder)	Hidden layers: {5, 7, 9}; hidden layer design (e.g., for 7 layers): $m \times (n_1, f{\cdot}n_2, g{\cdot}n_3, g{\cdot}n_4, g{\cdot}n_3, f{\cdot}n_2, n_1)$, where n_i is an integer between 1 to 5 with the condition that $n_j \geq n_i$ if $j > i$, m is the number of features, and f and g are fractions within {1/4, 1/2, 1/3, 2/3} with the condition that $f > g$. Activation function: {linear, logistic, tanh, relu}. Regularization: {0.00001, 0.0001, 0.001}. Early stopping: {True, False}
kNN	k: 1 to 23 by 1; distance weighting: uniform, inverse distance
LR	C: from 1 to 20 by 0.1; regularization penalty: L_1, L_2, elastic net, none
SVM	Kernel: linear, RBF, polynomial; C: {1, 10, 100, 1000, 10000}; γ: 1, 2, auto-scaled by the inverse of number of features and feature variance
GP	Kernel: {dot product, RBF}; length scale: {0.1, 1.0, 10.0}
RF	Number of estimators: {500, 1000, 2000}; splitting metric: {Gini, entropy}; maximum depth: {3 to 30}; maximum features: {1 to the number of autoencoder features used}; minimum number of samples for a split: {2 to 20 by 1}; minimum number of samples per leaf: {2 to 20 by 1}
ET	Same as for RF
AB	Same as for RF
ANN	Hidden layers: {3, 4, 5}; hidden layer design: (m,…, m, m/2), where m is the number of features and is repeated to reach the number of hidden layers; all other parameter ranges same as for ANN (autoencoder)

on the basis of their ability to separate prospective and non-prospective sites, regardless of the details of the workflow. However, the classification-threshold invariance can lead to issues with the usability of prospectivity maps because the investment cost associated with false negatives is zero, compared with the non-zero investment cost associated with false positives. Because we controlled class imbalance through purposely constructing two identically sized positive and negative label sets, the effect of class imbalance is minimized for all performance metrics. A summary for the MPM workflow, particularly as it pertains to the experimental controls and variables within the workflow, is shown in Figure 3.

RESULTS

Feature Extraction

The optimized autoencoder within the grid range has the structure (102, 51, 34, 25, nodes in the coding layer, 25, 34, 51, 102) (Fig. 4) with early stopping enabled, a regularization strength of 0.0001 and using the tanh activation function. The coding layer size can be controlled to modulate the performance of the autoencoder. Varying the coding layer size from 25 to 5 for a total of 20 runs (with randomized initializations of the nodal weights) per size reveals a gradual loss of autoencoder performance (Fig. 5). This loss has two distinguishable regimes, with a shallow trend from 25 to 10 and a steeper trend from 10 to 5 (Fig. 5). The reconstruction performance is often above a coefficient of determination (CoD) value of 0.95 for coding layer sizes above 11, which implies that less than 5% of the variability of the data was lost. This implies that the evidence layers are highly compressible and about 11 fully de-correlated latent features contain about 95% of the original variability. It is important to recognize that the rate of information loss (determined through the CoD metric) is significantly slower than the rate of dimensionality loss, especially down to a coding layer size of 8 (see shallow slope of curve, Fig. 5). Consequently, the main difference imparted onto downstream algorithms is more attributable to changes in feature space dimensionality, rather than feature information-content.

To provide the readers with a qualitative visualization of the latent data, we visualized the extracted features, for two features for coding layer sizes 5 and 25 (Fig. 6). There was excellent spatial variability that roughly followed geological terranes (e.g., around the Canadian Shield, across the interior plains and the Cordilleran range). However, the exact nature of the features was not geologically interpretable because they were complex nonlinear combinations and transformations of the input data. Moreover, it was theoretically clear and empirically observable (Fig. 6) that nonlinear combinations resulted in features exhibiting nonlinear spatial variability. This may also be the case for many evidence

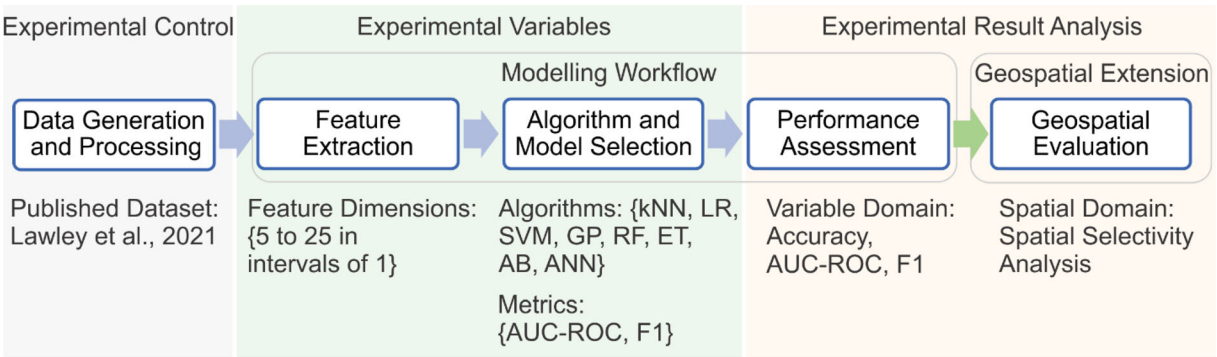


Figure 3. Illustration of the experimental MPM workflow, depicting the experimental controls and variables, as well as the result analysis methods. The experimental variables are modulated, whose choices are provided in set notation (members in curly brackets).

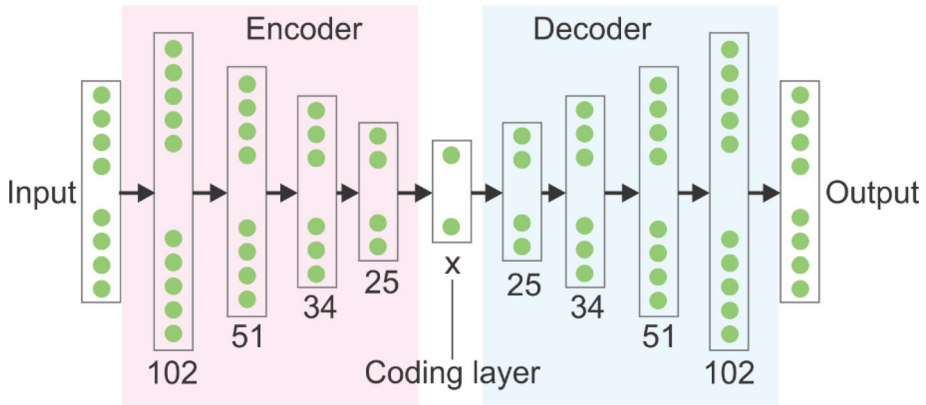


Figure 4. Network configuration for the optimized autoencoder. The coding layer size is shown as "*x*". The black arrows abbreviate full connectivity between neurons in adjacent layers (fully connected layers).

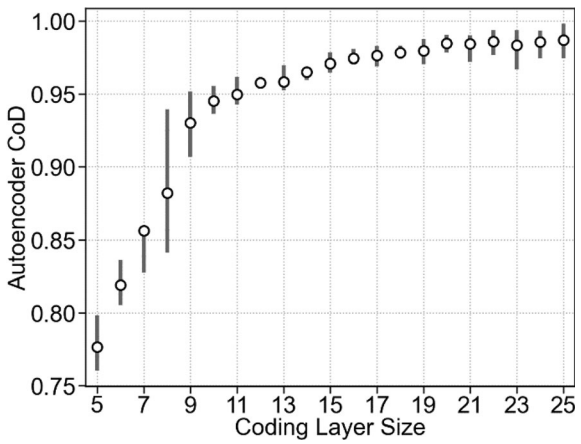


Figure 5. Autoencoder's performance as measured through the coefficient of determination (CoD) metric, as a function of the coding layer size of the autoencoder.

layers even prior to encoding because many types of data are discontinuous across terranes, faults or other types of boundaries (e.g., geochronological age and rock type). Incidentally, this implies that linear geostatistical methods (including interpolation) are not generally reliable on nonlinear latent data.

Predictive Modeling

For each set of autoencoder features from 5 to 25, we performed predictive modeling using all eight machine learning algorithms (Table 1), with the model selection driven by either the AUC-ROC or weighted F1 metric. This resulted in two sets of 168 optimal models with a range of performance (Figs. 7

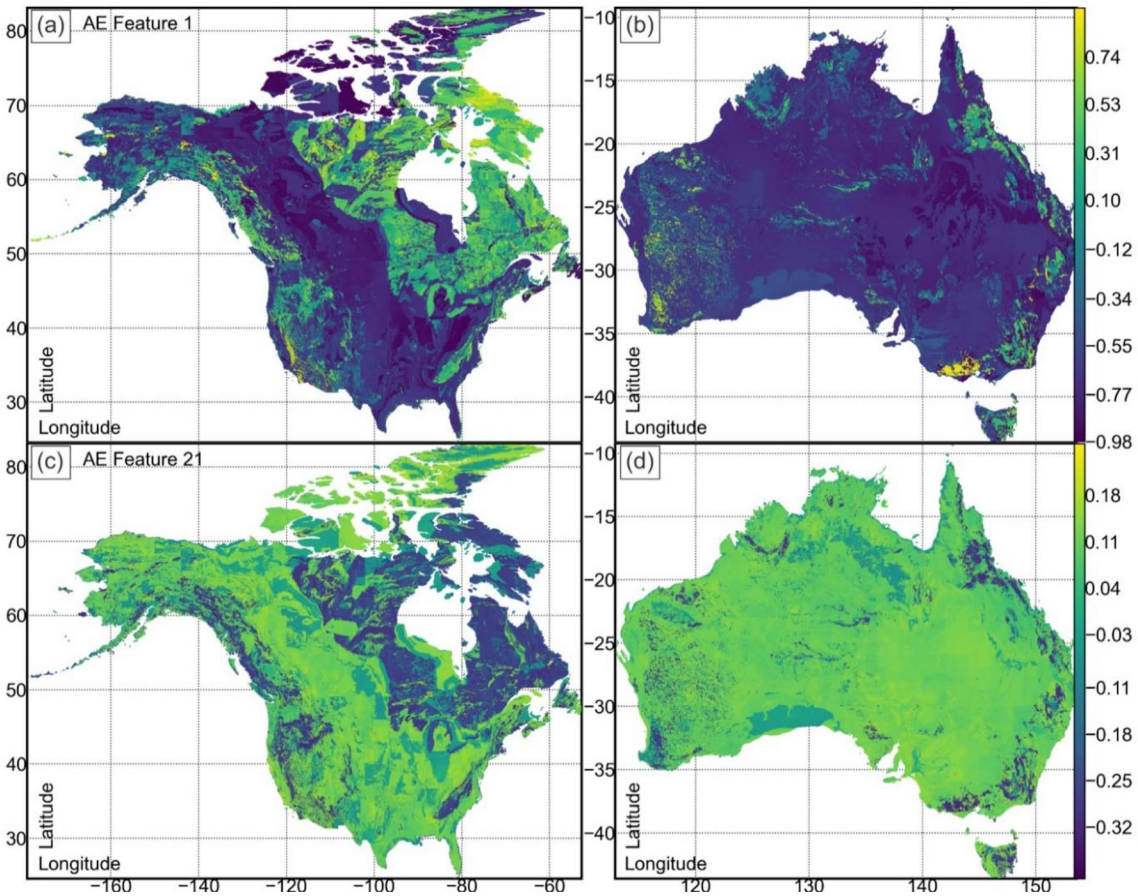


Figure 6. Visualization of the autoencoder's extracted features (AE features) 1 in (**a**) and (**b**), and 21 in (**c**) and (**d**) for coding layer sizes 5 and 25.

and 8). Two categories of algorithms are evident based on the results—those that were performance-sensitive or -insensitive to the number of features used. The first category consisted of SVM, LR, ANN and GP, all of which generally exhibited an increase in all metric scores up to about 20 features or coding layer nodes (Figs. 7b, f, g, and h and 8b, f, g, and h). Beyond a coding layer size of 20 nodes, model performance in this category demonstrated plateaus or slight decreases for models that were selected using the AUC-ROC metric (Fig. 7). For models selected using the weighted F1 metric, the trend was somewhat similar but the transition was not as clear for some algorithms, such as the ANN algorithm, which demonstrated a performance plateau past 20 nodes (Fig. 8g). For other algorithms, the performance plateau was not reached until 21 or 22 (e.g., AB and SVM; Fig. 8e and b, respectively) or slightly more (e.g., ET and GP; Fig. 8d and h). The second category consisted of kNN, RF, ET and AB. This category tended to exhibit the same model performance with minor variance (Fig. 7a, c, d, and e). This is an important observation because, clearly, the strategy to perform dimensionality reduction depends on the choice of algorithms and to a lesser extent, the choice of performance metric for performance assessment. An additional observation was that hyperparameter tuning using the AUC-ROC metric increased the AUC-ROC score of all best models for each algorithm, whereas tuning using the weighted F1 metric resulted in generally lower performance metric scores for all performance metrics used (Figs. 7 and 8). It is also worth to note that hyperparameter tuning using the weighted F1 metric in our study resulted in essentially identical metric scores for all metrics used, but this would not be expected where class imbalance existed.

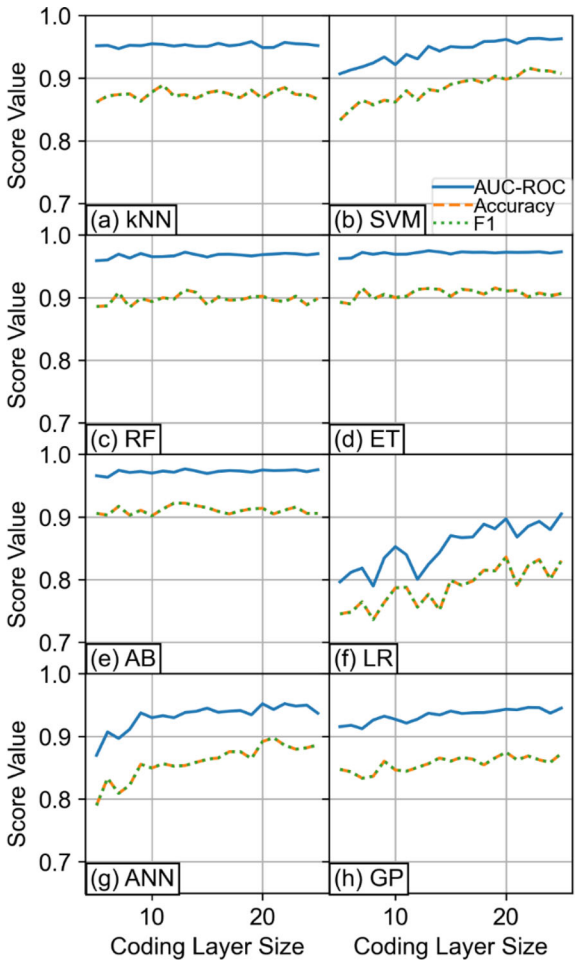


Figure 7. Performance of various algorithms with hyperparameter tuned using the AUC-ROC metric during predictive modeling as a function of the number of autoencoder features used (as controlled by the coding layer size).

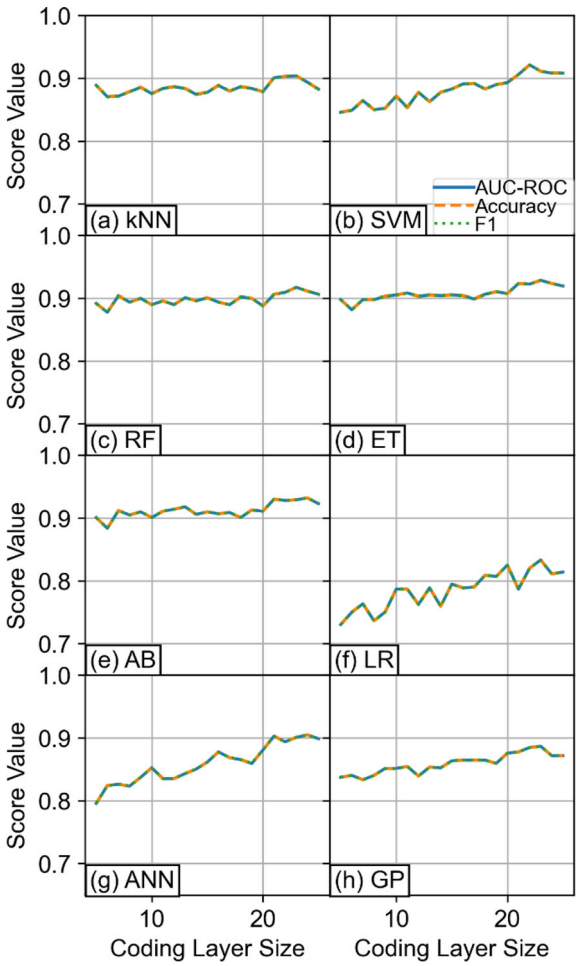


Figure 8. Performance of various algorithms with hyperparameter tuned using the weighted F1 metric during predictive modeling as a function of the number of autoencoder features used (as controlled by the coding layer size).

Analyzing prospectivity maps in solely the spatial domain is challenging to standardize because there are no data science metrics for this task, and it is impossible to craft universal metrics because there are no universal notions of spatial quality. The closest metric is related to the prediction–area curve, which simultaneously assess both spatial and variable domain characteristics (Yousefi & Carranza, 2015a, 2015b). The ratio of class-dependent prediction rate (sites per class divided by total number of sites) to its corresponding occupied area (area of the class divided by total area) is known as normalized density (Mihalasky & Bonham-Carter, 2001). Consequently, normalized density is a composite metric that is formulated like a posterior probability with the numerator incorporating variable domain performance, and the denominator spatial selectivity of models. Prediction rate is somewhat comparable, but not identical to the accuracy metric, and is incomparable to either the F1 or AUC-ROC metrics. This is a problem for our intended usage because we cannot explicitly and generally relate model quality in terms of purely data science metrics, which are always used to guide model construction, to model quality in terms of purely geospatial characteristics. Therefore, we extended the idea of the normalized density metric as a ratio of two independent metrics (across domains) to incorporate typical data science metrics, eschewing prediction rate. It was also unnecessary that the numerator–denominator assignment be kept the same as it was in normalized density. Spatial selec-

tivity was assessed through the occupied area metric, which was calculated in this study as the fraction of sites prospective (fraction prospective). Consequently, visualizing spatial selectivity as a function of metrics in the data science framework revealed the spatial evolution of models given workflow variability. Therefore, the spatial-variable domain coupling can be directly related to model tuning. An excessively overfitted model would predict few sites beyond positive training labels as prospective. Therefore, it was possible to compare, among other things, the fitting quality of models in response to changes in feature space dimensionality between algorithms, by examining the changes in spatial selectivity vs. model performance. A physical upper bound on spatial selectivity (occupied area) is unknown because the total area of mineral deposits cannot be known, but heuristic order-of-magnitude guesses may be possible (e.g., Fermi's piano tuner problem; Zipf's Law; Merriam et al., 2004).

In terms of spatial selectivity alone, tree-based methods (group 1) exhibited the highest selectivity to within the range of parameters studied, across all sizes of feature dimensions (Fig. 9). The moderately selective algorithms included the kNN, ANN, SVM and GP (group 2). However, the variance of selectivity within this group was relatively large compared to that of the tree-based group (Fig. 9). The least selective group (group 3) consisted of solely LR, which did not overlap with any other group at all sizes of feature dimensionality (Fig. 9). Within groups, there was a common trend of increasing spatial selectivity with increasing coding layer size (feature space dimensionality), although at feature space dimensions beyond about 17, there was a gradual plateauing and loss in selectivity for kNN, ET, AB and RF algorithms, which was prominent for models selected using the weighted F1 metric (Fig. 9b). The net change in selectivity (average slope) was not comparable for all groups. The average slopes for the tree-based algorithms were shallower than those of the other groups, with the exception of the kNN algorithm, which was comparable (Fig. 9). To provide a robust summary, we performed a sensitivity analysis of spatial variability induced by changes in workflow, which we determined using the mean range metric (the mean of the absolute range of scores). The fraction prospective metric was the most sensitive to changes in algorithm, which resulted in a mean range of 0.15 across all feature space dimensions and model selection metrics. The second-most sensitive variable was the

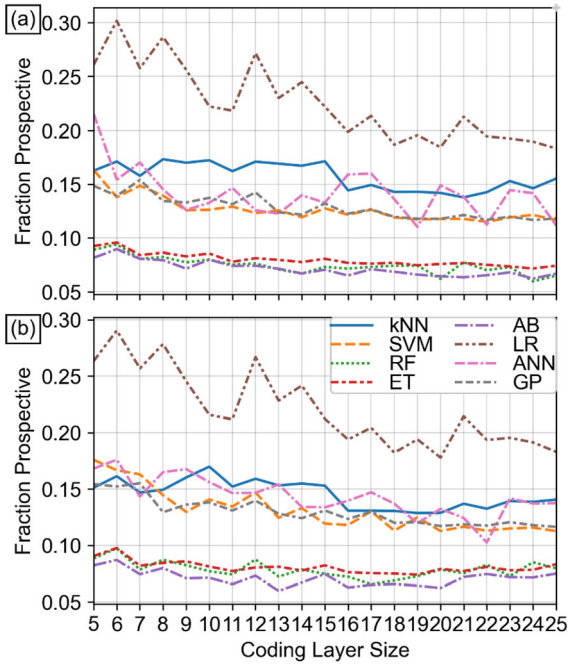


Figure 9. The fraction of sites that are deemed prospective (fraction prospective) as a function of the number of features used (as controlled by the coding layer size) for hyperparameter tuning using the AUC-ROC metric (**a**) and the weighted F1 metric (**b**).

feature space dimensionality (coding layer size), with a mean range of 0.05, across all algorithms and all model selection metrics. The least sensitive variable was the choice of tuning metric, with a mean range of 0.01, across all algorithms and feature space dimensions.

The relationships between spatial selectivity and model performance (the numerator and denominator of the composite metric) were noisy and varied (for the AUC-ROC metric, see Fig. 10, the results were qualitatively similar for the weighted F1 metric). There were two distinct categories: (1) those that exhibit clear relationships between spatial and variable domain characteristics, which included SVM, LR, ANN and GP (Fig. 10b, f, g, and h); and (2) those that exhibited no strong relationship between spatial and variable domain characteristics, which included kNN, RF, ET and AB (Fig. 10a, c, d, and e). In effect, only some algorithms resulted in models whose spatial selectivity was sensitive to variable domain performance (the "spatial sensitivity issue"). In addition, the tree-based algorithms (RF, ET and AB) produced models that exhibited an inflection in their variable–spatial domain relationships, which can be observed

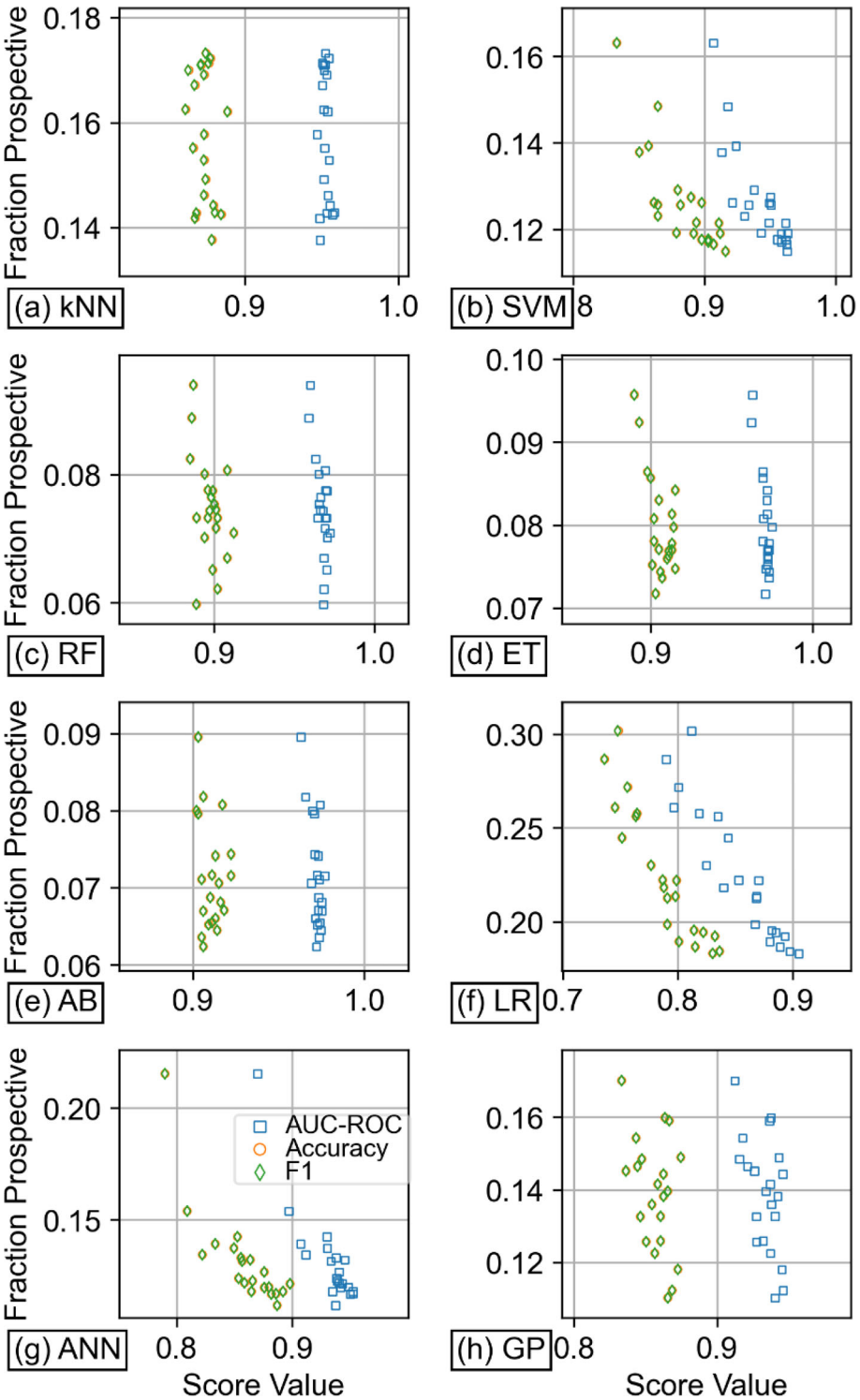


Figure 10. Relationships between spatial selectivity as measured by the fraction of sites predicted as prospective (fraction prospective) and variable domain metric scores.

as a curvature in the scatter plots (Fig. 10c, d, and e). The peak variable domain performance for the tree-based methods occurred at roughly the mid-range of spatial selectivity and away from the mid-range, the performance was generally lower as measured through any performance metric. Therefore, the relationship between variable domain performance and spatial selectivity was not monotonic and exhibited at least a peak (the ''selectivity ambiguity issue''). The existence of the peak is telltale of a gradual overfitting of the models that may not be detectable within heuristic designs of the workflow. Consequently, because of both sensitivity and ambiguity issues, variable domain metrics cannot generally guide model selection toward a target degree of spatial selectivity.

Because MPM results in maps that are often used visually, it is also important to examine qualitatively the resulting prospectivity maps. For each set of maps resulting from a unique combination of the number of features and model-tuning metric, there were observable differences between the best models for each algorithm (e.g., Fig. 11). The differences were the greatest between the best and worst performing models regardless of the performance metric used to tune the models (Fig. 9). To understand the degree of consensus within the two sets of 168 optimal models, we summarized the differences of each set using a linearly weighted average, based on the tuning metric used (AUC-ROC or F1 score). Weighted averaging seeks to reduce the disproportionate impact of the poorest performing models. This yields a set of maps that are not intended to be prospectivity maps per se (because this is not an applicative study), but are experimental 'consensus maps' (equivalent to a workflow uncertainty analysis). Values on the maps that are closer to 1 imply consensus for a positive label and at 0 consensus for a negative label, and in between meaning variable dissent or disagreement. To create a range in variable domain performance, we combined only models above a given performance level using their hyperparameter-tuning performance metric, producing three sets of maps with all models (cut-off threshold of 0.0) and only those above 0.85 and 0.90 (Figs. 12 and 13). These thresholds were not classification thresholds, but were performance criteria for model merging. It is visually clear that there was better consensus at higher model performance, which can be seen in the form of a loss of intermediate-colored (colors near the center of the color bar) sites and more bright (more yellow) sites, and dark (more deep blue) sites (compare Fig. 12e with Fig. 12a). Regardless of the performance threshold, it was also visually obvious that consensus was significant in key prospective zones, which were consistently colorized yellow (e.g., western Canada extending into Alaska; Fig. 12). Conversely, the consensus was poor, e.g., in western to the central United States (Fig. 12).

It is visually discernible that models tuned using the weighted F1 metric were more selective spatially (e.g., Fig. 13f compared with Fig. 12f). Indeed, the quantitative difference can be substantial in maps (Fig. 14). In fact, for most algorithms with the exception of GP and kNN, the general trend was such that hyperparameter tuning using the weighted F1 metric resulted in the most spatially selective models across a range of feature space dimensionality (Fig. 15). Spatial selectivity also increased more rapidly at increasing AUC-ROC thresholds for models that were tuned using weighted F1 metric than the AUC-ROC metric (Fig. 16). This observation combined with the observation that hyperparameter tuning using the weighted F1 metric resulted in generally lower AUC-ROC scores (Figs. 6 and 7), which implies that although there was a rough relationship between spatial selectivity and model performance metric scores on average, it varied depending on the performance metric used and is algorithm dependent (Fig. 16). A strategy to tune models to maximize AUC-ROC score is ineffective to create models with the highest spatial selectivity for our workflow and dataset.

DISCUSSION

In applicative domains of data science, model performance is usually stipulated by a client and through unlimited choices in workflow design, and the data scientist attempts to locate at least one such model satisfying the performance constraint. In this research study, we adopted a reasonable performance threshold of 0.90 as assessed through the AUC-ROC metric. Models meeting this criterion were considered satisfactory in the data science framework. Results suggest that there always exist at least one combination of feature space density/dimensionality and algorithm that produces an AUC-ROC score above 0.90 (Fig. 7). This is false for models produced using the weighted F1 metric because all LR and GP models fail this criterion (Fig. 8). Consequently, we demonstrated that it is

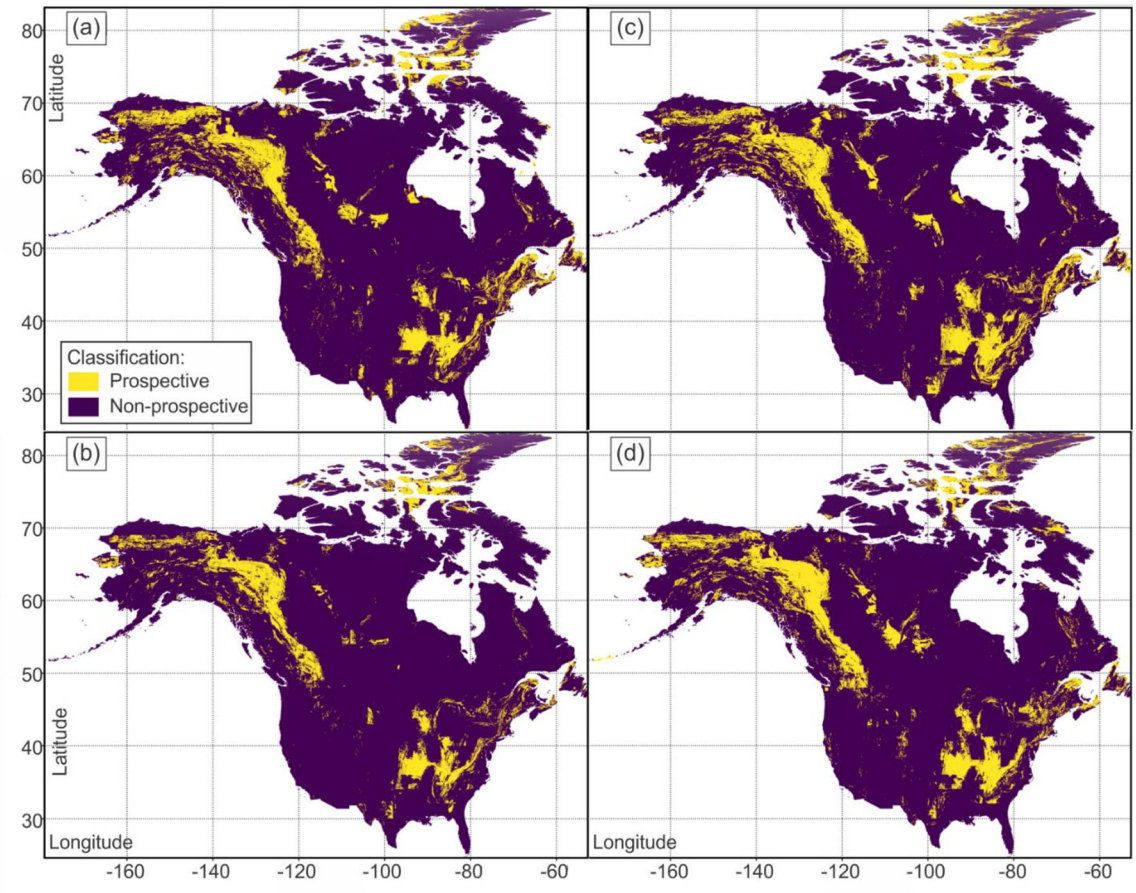


Figure 11. Visual comparison of the prospectivity maps produced using 19 features and the AUC-ROC metric. Results are shown for ANN in (**a**), RF in (**b**), GP in (**c**) and SVM in (**d**).

highly possible for MPM workflows to differ in key stages, including feature selection (or extraction using deep learning), predictive modeling algorithms and performance metrics, while still producing equiprobable models (in the case of classification tasks). The effect of feature space dimensionality was also significant. The increase in spatial selectivity of all algorithms with increasing feature space dimensionality, at least up to 20 features was expected given the increase in explanatory power of the features (Fig. 5). However, beyond 20 to 22 features, the curse of dimensionality effect likely caused performance plateaus or slight losses in model performance (Figs. 7 and 8). This implies that feature space density is becoming too low, such that, despite increases in the explanatory power of the features, there is a loss in model performance because of increasing sparsity of training samples relative to feature space dimensionality. In general, there was a noticeable to significant difference in prospectivity maps both quantitatively and qualitatively as a function of changes in algorithm and changes in feature space dimensionality. An implication of this finding is that feature engineering and algorithm interactions should be experimentally modified using feedback as part of the workflow, in the sense that a range of feature space dimensions should be explored for each algorithm to optimize model performance. Another implication of this finding is that, where MPM adopts machine learning, a comprehensive implementation of a data science framework should be a minimal requirement for model deployment settings (e.g., Fig. 1). This implication may also apply to data pre-processing, which could include the construction of evidence layers using discipline-specific methods, such as geostatistics, because clearly differences in data preprocessing would also propagate to machine learn-

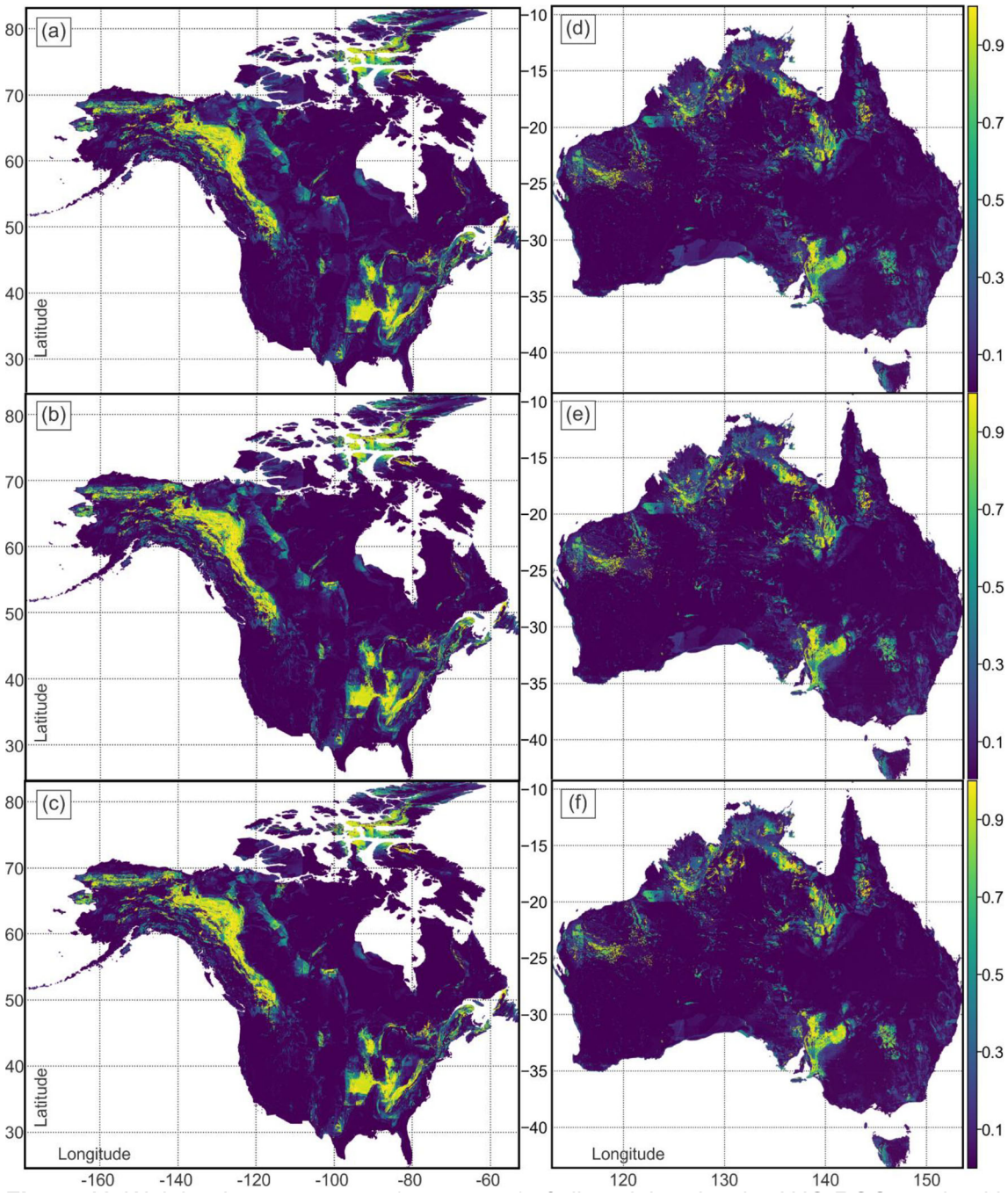


Figure 12. Weighted average maps (consensus) of all models using the AUC-ROC metric with a performance score: > 0.0 (**a**, **d**); > 0.85 (**b**, **e**); > 0.90 (**c**, **f**).

ing features. However, this exploration was outside of the scope of this study.

Another major finding was that there was not always a sensitive or monotonic relationship between spatial selectivity and model performance as assessed through the AUC-ROC, F1 or accuracy metrics. This is a potential problem because the geospatial extension to the data science framework

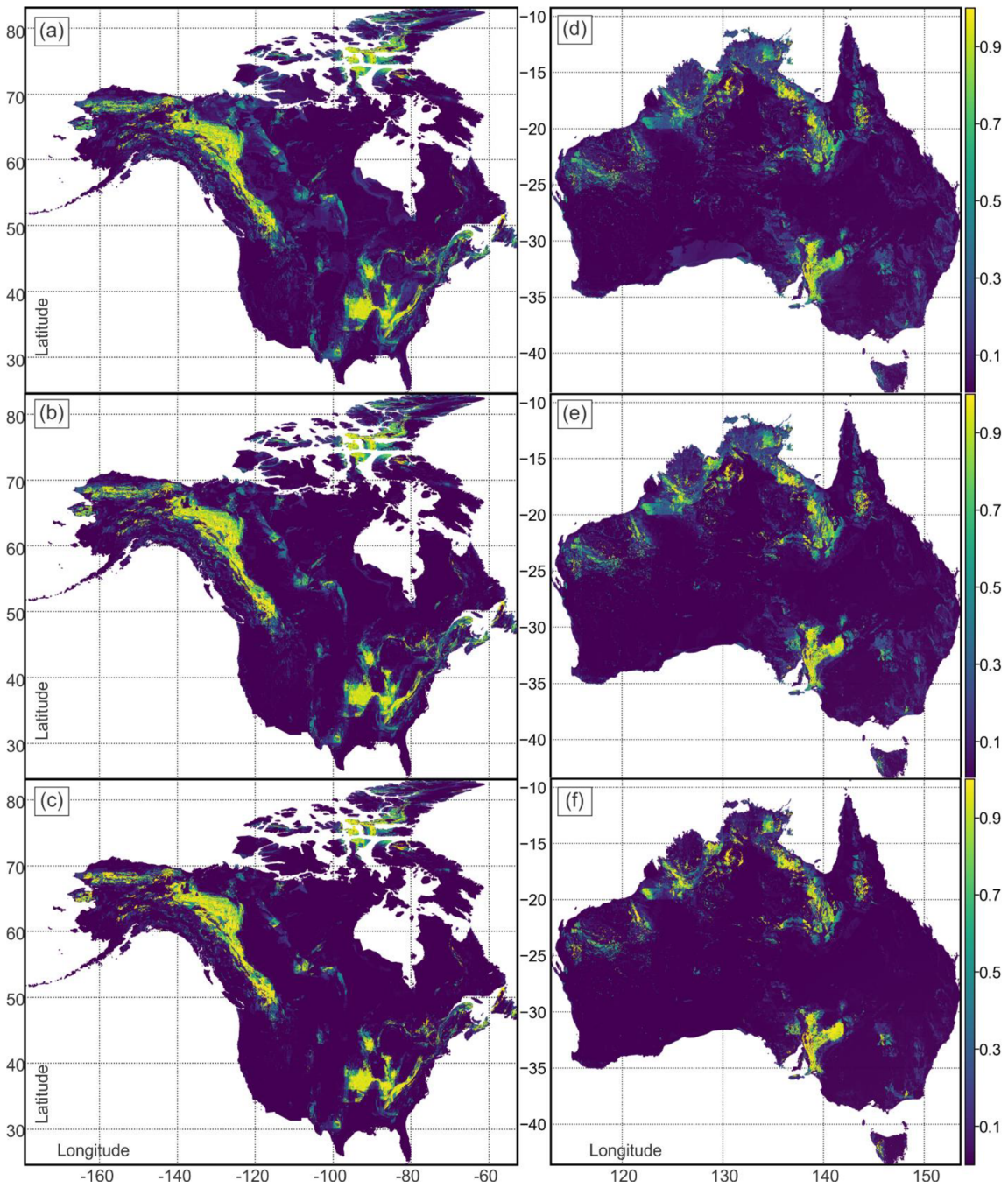


Figure 13. Weighted average maps (consensus) of all models using scores of the F1 metric with a performance score: > 0.0 (**a**, **d**); > 0.85 (**b**, **e**); > 0.90 (**c**, **f**).

is a post-hoc one. Therefore, e.g., it is unknown how spatially-guided model construction (e.g., hyperparameter tuning) could be integrated into the data science framework in general (i.e., not necessitating spatially aware algorithms). In our case, for the explored tree-based algorithms, which are popular in current MPM practices, optimizing model performance via any metric did not necessarily result in more spatially selective maps. On average, optimizing the F1 score, by using the F1 metric to per-

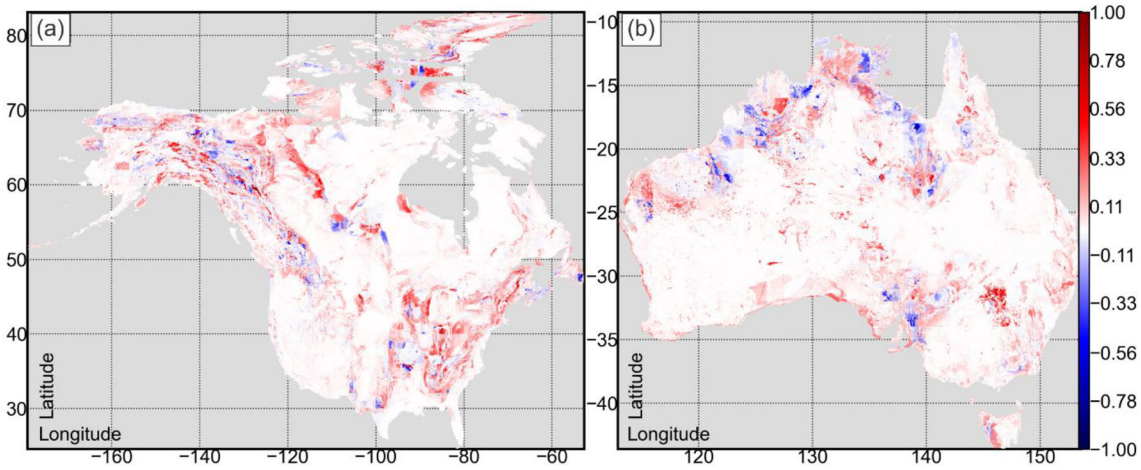


Figure 14. Differences in the consensus maps produced by tuning hyperparameters using the AUC-ROC metric vs. the weighted F1 metric (using only models with metric scores above 0.90). The difference is calculated as the consensus map produced using the AUC-ROC metric minus that of the F1 metric.

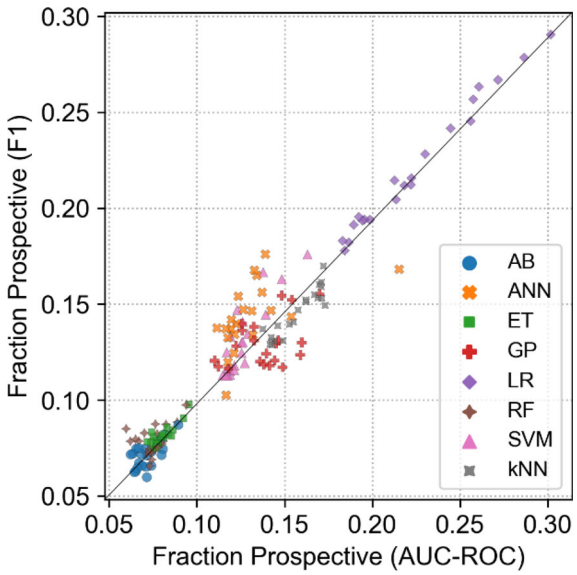


Figure 15. Fraction of all maps that are deemed prospective of all models tuned using the weighted F1 metric vs. those tuned using the AUC-ROC metric.

form model selection through hyperparameter tuning, produced more spatially selective maps. However, this produced models with generally lower AUC-ROC (as well as accuracy and F1) scores. This implies that the choice of metric during hyperparameter tuning was a significant source of variability in prospectivity maps in the spatial domain. Because the overarching objective of MPM is to narrow the search space for mineral resources, this implies that the value of MPM products cannot be determined within the data science framework because this framework employs solely variable domain metrics to select algorithms and models. Consequently, the rigor in which the geospatial extension (Fig. 1) is practiced significantly controls the value of MPM products. However, it is not obvious whether the extension of the data science framework using a geospatial post-hoc analysis can create effective feedback, such that the metric-driven aspect of the data science framework is preserved. For example, metrics in the spatial domain relying on large portions of maps could violate the train-test independence of datasets in data science. As such, it is not feasible to treat the style of geospatial extension of the data science framework as practiced in this study for the task of MPM, as a rigorous geodata science framework because cyclical tasks, such as data preparation to modeling cannot be generally conducted using metric-driven feedback and experimentation. For solely prospectivity mapping using this type of framework, some measure of spatial concordance between prospective sites and actual mineral occurrences would be an effective validation of any MPM product. However, because for many commodities or deposit types, the positive label is a constraint on training data abundance, it is not practical in general to set aside a spatially extensive and statistically large number of labels for robust model testing.

Unfortunately, the inconsistent and sometimes unreliable relationship between performance of models in the variable and spatial domains implies that out-of-sample testing and metric-driven model

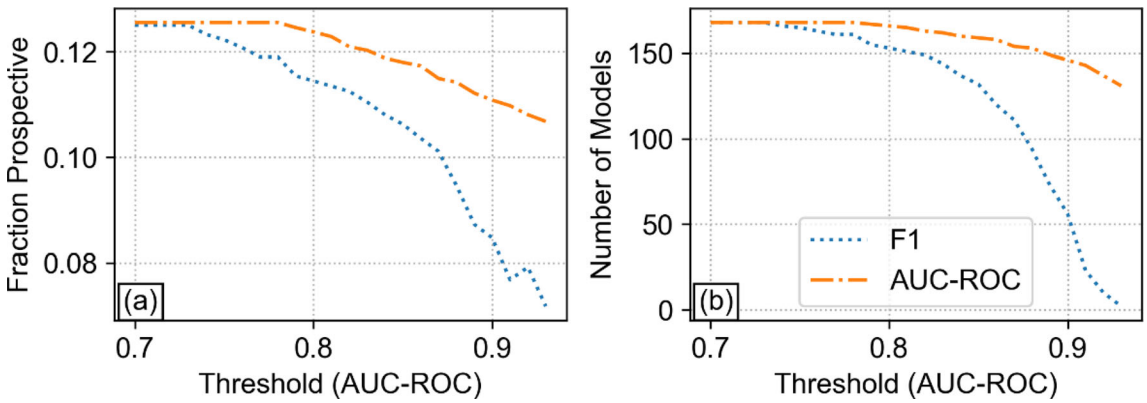


Figure 16. Comparison of the fraction of sites classified as prospective, averaged over all models meeting the performance threshold in (**a**) and the number of models satisfying the performance threshold in (**b**), both as a function of threshold values of AUC-ROC.

selection are insufficient to control models' spatial selectivity, the (unconstrained) optimization of which is the fundamental goal of MPM. This is demonstrably true for supervised classification approaches using spectrum-based (aspatial) algorithms and we expect that it would hold true also for regression-based MPM, but it remains to be empirically examined. This is likely more problematic for singleton or experiential studies (studies that result in single models or are based on subjective experience). This further implies that the uncertainty of MPM products is generally unknown in the sense of their spatial characteristics, e.g., how to define a desirable extent of reduction in exploration search space, and how to relate it to the data science framework. With the current state of theoretical development in the geodata science framework, it is not yet possible to design an objective workflow and its constraints (e.g., of number of features, algorithms and otherwise) to definitively produce a MPM product meeting or exceeding spatial requirements. Physical feedback approaches are not scalable in general to MPM products. For example, in mineral resource estimation, additional sampling could be used to delineate between resources and reserves, which could occur after the resource modeling stage. For MPM, this cannot generally happen to the scale necessary to validate maps. Therefore, feed-forward approaches are probably the only viable solution for now. Our results indicate that a possible way to de-risk MPM products is via consensus of the outputs of multiple workflows, whose key experimental variables are as de-correlated as possible, as was utilized in this study (Fig. 17). Any workflow component could be varied,

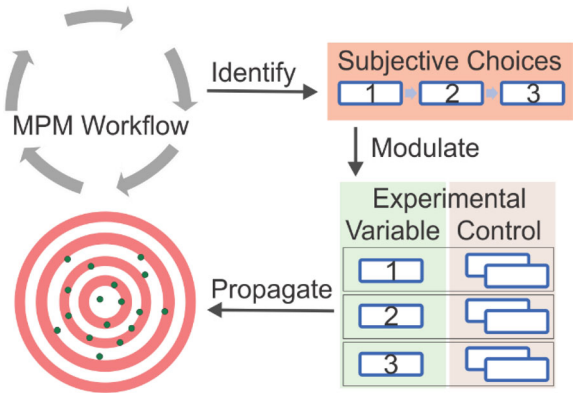


Figure 17. Summary of uncertainty propagation through workflow modulation to create consensus of mineral prospectivity mapping (MPM) products. The key stages depicted are identification of subjective choices, modulation of experimental variables and controls among the choices, and finally, propagation of uncertainty to determine degree of consensus (e.g., the uncertainty) of MPM products. A consensual model can then be constructed using many individual models to understand exploration risks created by workflow design.

beyond those examined in this study (e.g., the interpolation of sparse data into evidence layers). As such, we propose that a practical mitigation measure to workflow-induced variability is to build a large variety of experimental models, each of which is optimized through a single variation of a data science workflow (e.g., the highest scoring model under a set of experimental controls and variables, Fig. 3), from which, a consensual model is derived that depicts regions of varying consensus. In this way, exploration risk can be stratified (categorized) based on the degree of consensus or a level of tolerance of dissent. The most optimal exploration model is then

the subset of regions depicted by the consensual model that meets or exceeds a target degree of risk, which is defined by users of MPM products (e.g., 99% consensus or 1% dissent). Unfortunately, this means that prospectivity maps produced using singular workflows should probably be better treated as isolated experimental findings, in the sense that they require multiple either ensemble-type or subsequent replications via de-correlated confirmation studies or workflows to construct a scientific consensus. This would make sense given that MPM is technically a type of data-driven experimentation. Where consensus is likely to be a deciding factor for downstream investment are greenfield settings, where no additional knowledge would be available to constrain data-driven findings. To quantify uncertainty, it would only require modulation of the workflow to produce a set of (equiprobable) models or maps. To mitigate uncertainty, a threshold could be used to eliminate prospective areas of high disagreement from further consideration.

CONCLUSIONS

Uncertainty of MPM products derived from data-driven methods can be traced to the guiding frameworks. For MPM methods that are formulated as artificial intelligence tasks, the data science framework is the most common framework to guide model construction. The decoupling of data science practices, such as the experimentation of algorithms and models, and geoscience needs, such as spatial characteristics of models, prevents effective feedback to guide objective and extensive experimentation within data-driven MPM workflows. However, it is still a matter of research how spatial constraints can be integrated into the data science framework in a manner that respects the automatability, objectivity and rigor of the framework. Therefore, although the data science framework was originally designed to standardize inferential data modeling, to remove operator subjectivity and increase replicability, the extension of data science frameworks via post-hoc additions of geospatial analysis does not preserve this intention because it cannot be used to steer workflow design. This is the context in which heuristic choices in workflow design creates unknown and hidden uncertainty, which is not addressed by the data science framework, and whose closest concept is judgment-related uncertainty in GIS-based MPM. Given heuristic workflow design, variability in MPM products is significant and individual MPM outcomes (e.g., models or maps) contain unknown uncertainty with respect to choices in workflow design.

This study contributes to the understanding of uncertainty in MPM products. Specifically, we empirically demonstrated the existence of workflow-induced uncertainty for MPM workflows adhering to the geodata science framework. Moreover, we showed that its spatial effects are significant and unintuitive and exhibits complex relationships with model performance metrics. Specifically, we found that: (1) it is possible to reach high model performance through distinct combinations of feature-algorithm pairings, implying that there are many local minima in workflow design; (2) feature space dimensionality impacts model performance significantly, whose degree is algorithm dependent; (3) the choice of algorithms impact model performance but model performance alone is not sufficient to distinguish the quality of resulting maps as measured by spatial selectivity; and (4) there is an inconsistent relationship between model performance and spatial selectivity, which implies that MPM maps should not be solely derived using a data-driven methodology (e.g., the data science framework), but must involve some spatial or knowledge constraints.

We also identified a solution to minimize workflow-induced uncertainty, which leverages the notion of scientific consensus, reframing data science-based MPM into the greater experimental science framework. We demonstrated that consensus of workflows is realizable by treating the output of each workflow variation as an independent experimental measurement or outcome, then analyzing the degree of consensus across outcomes. Therefore, by carefully choosing experimental controls and variables, it is possible to both isolate uncertainty contributions associated with a particular stage of the workflow and study its interactions with those of other stages. In effect, we demonstrated that modulation of workflow design enables uncertainty propagation to the resulting MPM products. We expect our findings and solution to be generalizable to most known types of data science-based MPM because our assumptions are unspecific to our data or algorithms and we purposely explored a large (multi-continental scale) dataset and range of workflow implementations. We only anticipate that to adopt our solution, the key requirement would be computational capacity and high-performance computing expertise, as modulating workflows is com-

putationally intensive, even if it is fully automated. However, uncertainty analysis does not necessitate that all sources of variability to be propagated, only those with a dominant impact. Finally, in a philosophical interpretation, because data-driven MPM is a type of data-driven experimentation, it would be consistent with experimental science practices to ascribe value to outcome using scientific consensus. An implication of this finding is that where geodata science is used to derive scientific outcome, the geodata science framework should be embedded into an experimental science framework to properly assess the value of data-driven products.

ACKNOWLEDGMENT

The authors would like to thank Dr. Mohammad Parsa (Geological Survey of Canada) for discussions, which have improved the scope of this study. The authors thank three anonymous reviewers for their constructive comments, which have greatly improved this manuscript. We are also very thankful to Prof. E.J.M. Carranza and an anonymous Associate Editor for editorial handling.

FUNDING

Open access funding provided by Natural Resources Canada library. This research project was funded by the Critical Minerals Geoscience Data (CMGD) program at the Geological Survey of Canada (Natural Resources Canada).

DECLARATIONS

Conflict of 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.

OPEN ACCESS

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

REFERENCES

Abbott, D. (2014). *Applied predictive analytics*. Wiley.

Agterberg, F. P., & Bonham-Carter, G. F. (1999). Logistic regression and weights of evidence modeling in mineral exploration. In *Proceedings of the 28th international symposium on applications of computer in the mineral industry (APCOM), Golden* (vol. 483, pp. 490).

Alaska Division of Geological and Geophysical Surveys. (2016). Alaska merged geophysical data grids. *Alaska Division of Geological & Geophysical Surveys Data Series, 12*, 3. https://doi.org/10.14509/29555.

Arkani-Hamed, J. (2007). Differential reduction to the pole: Revisited. *Geophysics, 72*, L13–L20.

Bourdeau, J. E., Zhang, S. E., Lawley, C. J. M., Parsa, M., Nwaila, G. T., & Ghorbani, Y. (2023). Predictive geochemical exploration: Inferential generation of modern geochemical data, anomaly detection and application to northern Manitoba. *Natural Resources Research, 32*, 2355–2386.

Bradley, D. (2008). Passive margins through earth history. *Earth Science Reviews, 91*(1–4), 1–26.

Breiman, L. (1996a). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (1996b). Stacked regressions. *Machine Learning, 24*(1), 49–64.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Carranza, E. J. M., & Hale, M. (2001). Logistic regression for geologically constrained mapping of gold potential, Baguio district, Philippines. *Exploration and Mining Geology, 10*(3), 165–175.

Chen, Y. (2015). Mineral potential mapping with a restricted Boltzmann machine. *Ore Geology Reviews, 71*, 749–760.

Chen, Y., & Wu, W. (2017). Mapping mineral prospectivity using an extreme learning machine regression. *Ore Geology Reviews, 80*, 200–213.

Chen, Y., & Wu, W. (2019). Isolation forest as an alternative data-driven mineral prospectivity mapping method with a higher data-processing efficiency. *Natural Resources Research, 28*(1), 31–46.

Chen, Y., Wu, W., & Zhao, Q. (2020). A bat algorithm-based data-driven model for mineral prospectivity mapping. *Natural Resources Research, 29*(1), 247–265.

Chorlton, L. (2007). Generalized geology of the world: Bedrock domains and major faults in GIS format: A small scale world geology map with an extended geological attribute database. *Geological Survey of Canada, Open File, 5529*, 48. https://doi.org/10.4095/223767.

Coolbaugh, M. F., Raines, G. L., & Zehner, R. E. (2007). Assessment of exploration bias in data-driven predictive models and the estimation of undiscovered resources. *Natural Resources Research, 16*, 199–207.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions in Information Theory, 13*(1), 21–27.

Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute Working Paper No. 2002-119/4* (p. 16).

Curry, H. B. (1944). The method of steepest descent for non-linear Minimisation problems. *Quarterly of Applied Mathematics, 2*, 258–261.

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.

Debayle, E., Bodin, T., Durand, S., & Ricard, Y. (2020). Seismic evidence for partial melt below tectonic plates. *Nature, 586*, 555–563.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR), 27*(3), 326–327.

Ebbing, J., Haas, P., Ferraccioli, F., Pappa, F., Szwillus, W., & Bouman, J. (2018). Earth tectonics as seen by GOCE—Enhanced satellite gravity gradient imaging. *Scientific Reports, 8*, 16356.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fix, E., & Hodges, J. L. (1951). An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Review, 57*(3), 233–238.

Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23–37). Springer. https://doi.org/10.1007/3-540-59119-2_166.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139.

Geological Survey of Canada, 2017. *Geoscience data repository for geophysical data, gravity, point data.* Natural Resources Canada. Retrieved September 24, 2023, from https://geophysical-data.canada.ca/.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*, 3–42.

Gharehchopogh, F. S., Namazi, M., Ebrahimi, L., & Abdollahzadeh, B. (2023). Advances in sparrow search algorithm: A comprehensive survey. *Archives of Computational Methods in Engineering, 30*(1), 427–455.

Gondara, L. (2016). Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)* (pp. 241–246). IEEE. https://doi.org/10.1109/ICDMW.2016.0041.

Granitto, M., Giles, S., & Kelley, K. (2017). Global geochemical database for critical metals in black shales. *U.S. Geological Survey*, Data Release. https://doi.org/10.5066/F71G0K7X.

Grossi, V., Giannotti, F., Pedreschi, D., Manghi, P., Pagano, P., & Assante, M. (2021). Data science: A game changer for science and innovation. *International Journal of Data Science and Analytics, 11*, 263–278.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (vol. 2). Springer.

Hazzan, O., & Mike, K. (2023). The data science workflow. In *Guide to teaching data science*. Springer. https://doi.org/10.1007/978-3-031-24758-3_10.

Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E., & Smith, E. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). DAMA International.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd international conference on document analysis and recognition* (pp. 278–282). https://doi.org/10.1109/ICDAR.1995.598994.

Hoffimann, J., Zortea, M., de Carvalho, B., & Zadrozny, B. (2021). Geostatistical learning: Challenges and opportunities. *Frontiers in Applied Mathematics and Statistics, 7*, 689393.

Hoggard, M., Czarnota, K., Richards, F., Huston, D., Jaques, A., & Ghelichkhan, S. (2020). Global distribution of sediment-hosted metals controlled by craton edge stability. *Nature Geoscience, 13*, 504–510.

Horton, J., Juan, C., & Stoeser, D. (2017). The state geologic map compilation (SGMC) geodatabase of the conterminous United States (ver. 1.1, August 2017). *U.S. Geological Survey*, Data Series 1052, 46.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*, 457–506.

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: A review. *Complex & Intelligent Systems, 8*(3), 2663–2693.

Karbalaei Ramezanali, A., Feizi, F., Jafarirad, A., & Lotfi, M. (2020). Geochemical anomaly and mineral prospectivity mapping for vein-type copper mineralization, Kuhsiah-e-Urmak area, Iran: application of sequential Gaussian simulation and multivariate regression analysis. *Natural Resources Research, 29*(1), 41–70.

Kennett, B., Salmon, M., Saygin, E., & Group, A. W. (2011). AusMoho: The variation of Moho depth in Australia. *Geophysical Journal International, 187*(2), 946–958.

Kohavi, R., & Wolpert, D. H. (1996). Bias plus variance decomposition for zero-one loss functions. In *ICML* (Vol. 96, pp. 275–283).

Kotsiantis, S. B. (2014). Bagging and boosting variants for handling classifications problems: A survey. *The Knowledge Engineering Review, 29*(1), 78–100.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering, 160*(1), 3–24.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal, 37*(2), 233–243.

Lane, R., Wynne, P., Poudjom Djomani, Y., Stratford, W., Barretto, J., & Ceratori Tontini, F. (2020). 2019 Australian national gravity grids explanatory notes. *Geoscience Australia*, Record 2020/022. https://doi.org/10.11636/Record.2020.022.

Laske, G., Masters, G., Ma, Z., & Pasyanos, M. (2013). Update on CRUST1.0 - A 1-degree global model of Earth's crust. *EGU General Assembly*, EGU2013-2658.

Laudan, L. (1984). *Science and values: The aims of science and their role in scientific debate*. University of California Press.

Lawley, C., McCafferty, A., Graham, G., Gadd, M., Huston, D., Kelley, K., Paradis, S., Peter, J., & Czarnota, K. (2021). Datasets to support prospectivity modelling for sediment hosted Zn Pb mineral systems. *Geological Survey of Canada, Open File, 8836*, 13. https://doi.org/10.1016/j.oregeorev.2021.104635.

Lawley, C. J. M., McCafferty, A. E., Graham, G. E., Huston, D. L., Kelley, K. D., Czarnota, K., Paradis, S., Peter, J. M., Hayward, N., Bvarlow, M., Emsbo, P., Coyan, J., San Juna, C. A., & Gadd, M. G. (2022). Data-driven prospectivity modelling of sediment–hosted Zn–Pb mineral systems and their critical raw materials. *Ore Geology Reviews, 141*, 104635.

Lemaréchal, C. (2012). Cauchy and the gradient method. *Doc Math Extra, 251*(254), 10.

Li, Q., Chen, G., & Luo, L. (2023). Mineral prospectivity mapping using attention-based convolutional neural network. *Ore Geology Reviews, 156*, 105381.

Li, S., Chen, J., & Xiang, J. (2020). Applications of deep convolutional neural networks in prospecting prediction based on two-dimensional geological big data. *Neural Computing and Applications, 32*, 2037–2053.

Lin, N., Chen, Y., & Lu, L. (2020). Mineral potential mapping using a conjugate gradient logistic regression model. *Natural Resources Research, 29*, 173–188.

Luo, Z., Xiong, Y., & Zuo, R. (2020). Recognition of geochemical anomalies using a deep variational autoencoder network. *Applied Geochemistry, 122*, 104710.

Márquez, A. C. (2022). The curse of dimensionality. In *Digital Maintenance Management*. Springer Series in Reliability Engineering (pp. 67–86). Springer. https://doi.org/10.1007/978-3-030-97660-6_7.

McCafferty, A. E., San Juan, C. A., Lawley, C. J. M., Graham, G. E., Gadd, M. G., Huston, D. L., Kelley, K. D., Paradis, S., Peter, J. M., & Czarnota, K. (2023). National-scale geophysical, geologic, and mineral resource data and grids for the United States, Canada, and Australia: Data in support of the tri-national critical minerals mapping initiative: U.S. Geological Survey, Data release. https://doi.org/10.5066/P970GDD5.

Merriam, D. F., Drew, L. J., & Schuenemeyer, J. H. (2004). Zipf's law: A viable geological paradigm? *Natural Resources Research, 13*, 265–271.

Mihalasky, M. J., & Bonham-Carter, G. F. (2001). Lithodiversity and its spatial association with metallic mineral sites, Great Basin of Nevada. *Natural Resources Research, 10*, 209–226.

Miles, W., & Oneschuk, G. (2016). Magnetic anomaly map, Canada/Carte des anomalies magnétiques, Canada. *Geological Survey of Canada, Open File, 7799*, 1. https://doi.org/10.4095/297337.

Nwaila, G. T., Zhang, S. E., Bourdeau, J. E., Frimmel, H. E., & Ghorbani, Y. (2023). Spatial interpolation using machine learning: from patterns and regularities to block models. *Natural Resources Research, 33*, 129–162.

Nwaila, G. T., Zhang, S. E., Bourdeau, J. E., Ghorbani, Y., & Carranza, E. J. M. (2022). Artificial intelligence-based anomaly detection of the Assen iron deposit in South Africa using remote sensing data from the Landsat-8 operational land imager. *Artificial Intelligence in Geosciences, 3*, 71–85. https://doi.org/10.1016/j.aiig.2022.10.001.

OSC. (2016). National instrument 43-101: Standards of disclosure for mineral projects. *Ontario Securities Commission* (pp. 17).

Phillips, J. (2007). Geosoft eXecutables (GX's) developed by the U.S. Geological Survey, version 2.0, with notes on GX development from Fortran code. *U.S. Geological Survey*, Open-File Report 2007-1355 (p. 118).

Phillips, J., Duval, J., & Ambroziak, R. (1993). National geophysical data grids; gamma-ray, gravity, magnetic, and topographic data for the conterminous United States. *U.S. Geological Survey*, Data Series 9. https://doi.org/10.3133/ds9.

Porwal, A., Carranza, E. J. M., & Hale, M. (2004). A hybrid neuro-fuzzy model for mineral potential mapping. *Mathematical Geology, 36*, 803–826.

Porwal, A., Carranza, E. J. M., & Hale, M. (2006). Bayesian network classifiers for mineral potential mapping. *Computers & Geosciences, 32*(1), 1–16.

Porwal, A., Deb Das, R., Chaudhary, B., Gonzalez-Alvarez, I., & Kreuzer, O. (2015). Fuzzy inference systems for prospectivity modeling of mineral systems and a case-study for prospectivity mapping of surficial Uranium in Yeelirrie Area, Western Australia. *Ore Geology Reviews, 71*, 839–852.

Priestley, K., McKenzie, D., & Ho, T. (2018). A lithosphere–asthenosphere boundary—a global model derived from multimode surface-wave tomography and petrology. In H. Yuan & B. Romanowicz (Eds.), *Lithospheric Discontinuities* (pp. 111–123). American Geophysical Union. https://doi.org/10.1002/9781119249740.ch6.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning* (Vol. 1). MIT Press.

Ravat, D., Finn, C., Hill, P., Kucks, R., Phillips, J., Blakely, R., Bouligand, C., Sabaka, T., Elshayat, A., Aref, A., & Elawadi, E. (2009). A preliminary, full spectrum, magnetic anomaly grid of the United States with improved long wavelengths for studying continental dynamics: A website for distribution of data. *U.S. Geological Survey, Open-File Rep., 1258*, 7. https://doi.org/10.3133/ofr20091258.

Raymond, O., Liu, S., Gallagher, R., Highet, L., & Zhang, W. (2012). Surface Geology of Australia 1:1 000 000 scale 2012 edition. *Geoscience Australia, Commonwealth of Australia*, Canberra digital dataset. https://doi.org/10.26186/74619.

Reed, J. J., Wheeler, J., & Tucholke, B. (2005). Geologic map of North America—Perspectives and explanation. In *Decade of North America geology geologic map of North America—Perspectives and explanation* (vol. 1 pp. 1–28). Geological Society of America. https://doi.org/10.1130/DNAG-CSMS-v1.1.

Reguzzoni, M., & Sampietro, D. (2015). GEMMA: An Earth crustal model based on GOCE satellite data. *International Journal of Applied Earth Observation and Geoinformation, 35*(A), 31–43.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books. https://doi.org/10.1007/978-3-642-70911-1_20.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. Institute for cognitive science, University of California, ICS Report 8506 (p. 49).

Saad, O. M., & Chen, Y. (2020). Deep denoising autoencoder for seismic random noise attenuation. *Geophysics, 85*(4), V367–V376.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1249.

Saltus, R., Brown II, P., Morin, R., & Hill, P. (2008). 2006 compilation of Alaska gravity data and historical reports. *U.S. Geological Survey*, Data Series 264.

SAMREC (2016). *The South African code for the reporting of exploration results, mineral resources and mineral reserves (The SAMREC code)*. SAMREC 2016 edition (p. 98).

Schetselaar, E., & Snyder, D. (2017). National database of Moho depth estimates from seismic refraction and teleseismic surveys. *Geological Survey of Canada, Open File, 8243*, 14. https://doi.org/10.4095/305396.

Scotese, C. (2021). An atlas of Phanerozoic paleogeographic maps: The seas come in and the seas go out. *Annual Review of Earth and Planetary Sciences, 49*, 669–718.

Senanayake, I. P., Kiem, A. S., Hancock, G. R., Metelka, V., Folkes, C. B., Blevin, P. L., & Budd, A. R. (2023). A spatial data-driven approach for mineral prospectivity mapping. *Remote Sensing, 15*(16), 4074.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of data Warehousing, 5*(4), 13–22.

Shen, W., & Ritzwoller, M. (2016). Crustal and uppermost mantle structure beneath the United States. *Journal of Geophysical Research: Solid Earth, 121*(6), 4306–4342.

Shetty, S., Umesh, P., & Shetty, A. (2023). The effectiveness of machine learning-based multi-model ensemble predictions of CMIP6 in Western Ghats of India. *International Journal of Climatology*. https://doi.org/10.1002/joc.8131.

Shi, Z., Zuo, R., & Zhou, B. (2023). Deep reinforcement learning for mineral prospectivity mapping. *Mathematical Geosciences, 55*, 1–25.

Skabar, A. (2007). Mineral potential mapping using Bayesian learning for multilayer perceptrons. *Mathematical Geology, 39*, 439–451.

Snyder, D. B., Schetselaar, E., Pilkington, M., & Schaeffer, A. J. (2018). Resolution and uncertainty in lithospheric 3-D geological models. *Mineralogy and Petrology, 112*, 133–147.

Solomon, M. (1994). Social empiricism. *Noûs, 28*(3), 325–343. https://doi.org/10.2307/2216062.

Song, H., Gao, Y., Chen, W., Xue, Y. J., Zhang, H., & Zhang, X. (2020). Seismic random noise suppression using deep convolutional autoencoder neural network. *Journal of Applied Geophysics, 178*, 104071.

Song, S., Mukerji, T., & Hou, J. (2021). GANSim: Conditional facies simulation using an improved progressive growing of generative adversarial networks (GANs). *Mathematical Geosciences, 53*, 1413–1444.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*, 1929–1958.

Styron, R., & Pagani, M. (2020). The GEM global active faults database. *Earthquake Spectra, 36*(1), 160–180.

Sun, T., Chen, F., Zhong, L., Liu, W., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, Eastern China. *Ore Geology Reviews, 109*, 26–49.

Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z., & Hu, Z. (2020). Data-driven predictive modelling of mineral prospectivity using machine learning and deep learning methods: A case study from southern Jiangxi Province, China. *Minerals, 10*, 102.

Syberg, F. (1972). A Fourier method for the regional residual problem of potential fields. *Geophysical Prospecting, 20*(1), 47–75.

Szwillus, W., Afonso, J., Ebbing, J., & Mooney, W. (2019). Global crustal thickness and velocity structure from geostatistical analysis of seismic data. *Journal of Geophysical Research: Solid Earth, 124*(2), 1626–1652.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Proceedings of the USSR academy of sciences* (Vol. 39, pp. 195–198).

Wang, C., Chen, J., & Ouyang, Y. (2022). Determination of predictive variables in mineral prospectivity mapping using supervised and unsupervised methods. *Natural Resources Research, 31*, 2081–2102.

Wang, J., Zhou, Y., & Xiao, F. (2020). Identification of multi-element geochemical anomalies using unsupervised machine learning algorithms: A case study from Ag–Pb–Zn deposits in north-western Zhejiang, China. *Applied Geochemistry, 120*, 104679.

Wilson, F., Hults, C., Mull, C., & Karl, S. (2015). Geologic map of Alaska. *U.S. Geological Survey Scientific Investigations Map 3340*, 196, pp. 2. Retrieved September 24, 2023, from https://alaska.usgs.gov/science/geology/state_map/interactive_map/AKgeologic_map.html.

Xiong, Y., Zuo, R., & Carranza, E. J. M. (2018). Mapping mineral prospectivity through big data analytics and a deep learning algorithm. *Ore Geology Reviews, 102*, 811–817.

Yang, F., Wang, Z., Zuo, R., Sun, S., & Zhou, B. (2023). Quantification of uncertainty associated with evidence layers in mineral prospectivity mapping using direct sampling and convolutional neural network. *Natural Resources Research, 32*(1), 79–98.

Yang, N., Zhang, Z., Yang, J., & Hong, Z. (2022). Mineral prospectivity prediction by integration of convolutional autoencoder network and random forest. *Natural Resources Research, 31*(3), 1103–1119.

Yin, B., Zuo, R., & Sun, S. (2023). Mineral prospectivity mapping using deep self-attention model. *Natural Resources Research, 32*(1), 37–56.

Yin, J., & Li, N. (2022). Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping. *Ore Geology Reviews, 145*, 104916.

Yousefi, M., & Carranza, E. J. M. (2015a). Fuzzification of continuous-value spatial evidence for mineral prospectivity mapping. *Computers & Geosciences, 74*, 97–109.

Yousefi, M., & Carranza, E. J. M. (2015b). Prediction–area (P–A) plot and C-A fractal analysis to classify and evaluate evidential maps for mineral prospectivity modeling. *Computers & Geosciences, 79*, 69–81.

Yousefi, M., Carranza, E. J. M., Kreuzer, O. P., Nykänen, V., Hronsky, J. M., & Mihalasky, M. J. (2021). Data analysis methods for prospectivity modelling as applied to mineral exploration targeting: State-of-the-art and outlook. *Journal of Geochemical Exploration, 229*, 106839.

Yousefi, M., & Nykänen, V. (2016). Data-driven logistic-based weighting of geochemical and geological evidence layers in mineral prospectivity mapping. *Journal of Geochemical Exploration, 164*, 94–106.

Zhang, Q., Chen, J., Xu, H., Jia, Y., Chen, X., Jia, Z., & Liu, H. (2022a). Three-dimensional mineral prospectivity mapping by XGBoost modeling: A case study of the Lannigou gold deposit, China. *Natural Resources Research, 31*(3), 1135–1156.

Zhang, S. E., Bourdeau, J. E., Nwaila, G. T., & Corrigan, D. (2021). Towards a fully data-driven prospectivity mapping methodology: A case study of the Southeastern Churchill Province, Québec and Labrador. *Artificial Intelligence in Geosciences, 2*, 128–147.

Zhang, S. E., Bourdeau, J. E., Nwaila, G. T., & Ghorbani, Y. (2022b). Advanced geochemical exploration knowledge using machine learning: Prediction of unknown elemental concentrations and operational prioritization of re-analysis campaigns. *Artificial Intelligence in Geosciences, 3*, 86–100.

Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., Ghorbani, Y., & Carranza, E. J. M. (2023). Machine learning-based delineation of geodomain boundaries: A proof-of-concept study using data from the Witwatersrand goldfields. *Natural Resources Research, 32*(3), 879–900.

Zhang, Y., Li, A., & Hu, H. (2019). Crustal structure in Alaska from receiver function analysis. *Geophysical Research Letters, 46*(3), 1284–1292.

Zhang, Z., Wang, G., Carranza, E. J. M., Fan, J., Liu, X., Zhang, X., Dong, Y., Chang, X., & Sha, D. (2022c). An integrated framework for data-driven mineral prospectivity mapping using bagging-based positive-unlabeled learning and Bayesian cost-sensitive logistic regression. *Natural Resources Research, 31*(6), 3041–3060.

Zuo, R. (2020). Geodata science-based mineral prospectivity mapping: A review. *Natural Resources Research, 29*, 3415–3424.

Zuo, R., & Carranza, E. J. M. (2011). Support vector machine: A tool for mapping mineral prospectivity. *Computers & Geosciences, 37*(12), 1967–1975.

Zuo, R., Kreuzer, O. P., Wang, J., Xiong, Y., Zhang, Z., & Wang, Z. (2021). Uncertainties in GIS-based mineral prospectivity mapping: Key types, potential impacts and possible solutions. *Natural Resources Research, 30*, 3059–3079.

Zuo, R., Luo, Z., Xiong, Y., & Yin, B. (2022). A geologically constrained variational autoencoder for mineral prospectivity mapping. *Natural Resources Research, 31*(3), 1121–1133.

Zuo, R., Xiong, Y., Wang, J., & Carranza, E. J. M. (2019). Deep learning and its application in geochemical mapping. *Earth Science Reviews, 192*, 1–14.

Zuo, R., Xiong, Y., Wang, Z., Wang, J., & Kreuzer, O. P. (2023). A new generation of artificial intelligence algorithms for mineral prospectivity mapping. *Natural Resources Research, 32*, 1859–1869.

Zuo, R., & Xu, Y. (2023). Graph deep learning model for mapping mineral prospectivity. *Mathematical Geosciences, 55*(1), 1–21.

Zuo, R., & Xu, Y. (2024). A physically constrained hybrid deep learning model to mine a geochemical data cube in support of mineral exploration. *Computers & Geosciences, 182*, 105490.