Natural Resources Research, Vol. 33, No. 1, February 2024 (© 2023)
https://doi.org/10.1007/s11053-023-10280-7

Original Paper

Spatial Interpolation Using Machine Learning: From Patterns and Regularities to Block Models

Glen T. Nwaila,[1,6] Steven E. Zhang,[2] Julie E. Bourdeau,[2] Hartwig E. Frimmel,[3,4] and Yousef Ghorbani[5]

Received 14 May 2023; accepted 26 October 2023
Published online: 25 November 2023

In geospatial data interpolation, as in mapping, mineral resource estimation, modeling and numerical modeling in geosciences, kriging has been a central technique since the advent of geostatistics. Here, we introduce a new method for spatial interpolation in 2D and 3D using a block discretization technique (i.e., microblocking) using purely machine-learning algorithms and workflow design. This paper addresses the challenges of modeling spatial patterns and regularities in nature, and how different approaches have been used to cope with these challenges. We specifically explore the advantages and drawbacks of kriging while highlighting the long and complex sequence of procedures associated with block kriging. We argue that machine-learning techniques offer opportunities to simplify and streamline the process of mapping and mineral resource estimation, especially in cases of strong spatial relationships between sample location and resource concentration. To test the new method, synthetic 2D and 3D data were used for both 2D block modeling and geometallurgical modeling of a synthetic porphyry Cu deposit. The synthetic porphyry Cu data were very useful in validating the performance of the proposed microblocking technique as we were able to reproduce known values at unsampled locations. Our proposed method delivers the benefits of a machine learning-based block modeling approach, which includes its simplicity (a minimum of 2 hyperparameters), speed and familiarity to data scientists. This enables data scientists working on spatial data to employ workflows familiar to their training, to tackle problems that were previously solely in the domain of geoscience. In exchange, we expect that our method will be a gateway to attract more data scientist to become geodata scientists, benefitting the modern data-driven mineral value chain.

KEY WORDS: Geostatistics, Block model, Kriging, Mineral resources estimation, Machine learning, Patterns.

[1]Wits Mining Institute, University of the Witwatersrand, 1 Jan Smuts Ave., Johannesburg 2000, South Africa.
[2]Geological Survey of Canada, 601 Booth Street, Ottawa, ON K1A 0E8, Canada.
[3]Institute of Geography and Geology, Department of Geodynamics and Geomaterials Research, University of Würzburg, Am Hubland, 97074 Würzburg, Germany.
[4]Department of Geological Sciences, University of Cape Town, Rondebosch 7701, South Africa.
[5]School of Chemistry, University of Lincoln, Joseph Banks Laboratories, Green Lane, Lincoln, Lincolnshire LN6 7DL, UK.
[6]To whom correspondence should be addressed; e-mail: glen.nwaila@wits.ac.za

INTRODUCTION

Patterns and regularities are common phenomena in nature and represent a state of order and structure, and rational principles that govern nature (Good, 1983; Washburn et al., 1988; Dennett, 1991; Goertzel, 2006; Uddin & Hamiduzzaman, 2009; Kuipers, 2001; Steiner, 2009). They have been studied in a variety of fields, such as physics, mathematics, statistics, chemistry, biology, philosophy

1520-7439/24/0200-0129/0 © 2023 The Author(s)

and geosciences. In geosciences, spatial patterns are evident in the repetitive and consistent characteristics of rock types, mineral deposits and their formation processes (Groves et al., 2005). Spatial characteristics are of paramount importance in exploring for, and comprehending, the genesis of mineral deposits and can be leveraged to improve spatial data modeling from mapping to mineral resource estimation. The study of spatial patterns has been a focus of research in geosciences for hundreds of years and has led to the development of theories that propose that the repetitive and consistent characteristics of mineral deposits can be used to locate, decipher ore-forming processes and facilitate mineral resource estimations (Carranza, 2009).

In geosciences, spatial patterns are important to both knowledge and data-driven inquiries. Methods to model and study spatial patterns have resulted in the birth and development of many disciplines, such as geostatistics and remote sensing. Compared to traditional geostatistics, machine learning (ML), as a field, is more recent and presents pathways toward leveraging data that describe mineral deposits and their signatures (Samson, 2020; Dumakor-Dupey & Arya, 2021). ML algorithms have been employed in the study and exploration of mineral deposits to identify and model patterns and regularities that are unobvious (Srinivasan & Fisher, 1995; Galetakis et al., 2022; Mery & Marcotte, 2022). This has been demonstrated in studies such as applying ML algorithms to satellite imagery to locate and study mineral deposits, and to improve mineral exploration (Maxwell et al., 2018; Cevik et al., 2021; Diaz-Gonzalez et al., 2022; Liu et al., 2022; Nwaila et al., 2022). Particularly related to this study is the use of ML to perform geodomain boundary delineation (Zhang et al., 2023), which is a task that is required in the geostatistical treatment of mapping to resource estimation. Similarly, ML algorithms that could model spatial patterns are a powerful tool to improve or provide alternatives to geostatistical spatial modeling. The rigor of resource estimation (e.g., a need for reconciliation) implies that differences in spatial modeling are the most impactful and appreciable at this scale, although any benefits of a new approach would apply across all spatial scales.

In this paper, we propose a new method to perform spatial interpolation using ML algorithms in a manner similar to that of geostatistical block modeling. A key recognition in our method is that ML algorithms are generally unable to perform a change-of-support. Although this seems like a drawback of ML, we demonstrate that this property could be exploited, in combination of a re-examination of the assumption of support punctuality to perform a rigorous change-of-support. We compare and contrast the results using ML and typical geostatistical modeling using a large ensemble of synthetically generated datasets. We explore the differences between the traditional geostatistical approach and our proposed method in a variety of scenarios, using a combination of quantitative and qualitative comparisons through modulating a few key parameters: (1) sampling type – regular and biased random sampling; (2) sampling rate (both regular and biased random sampling); (3) strength of spatial correlation (through controlling the nugget effect); and (4) anisotropy. The results indicate that qualitatively and quantitatively, our method produces results that are competitive with block kriging across a wide range of conditions. A key benefit of our approach is that it requires a substantially reduced set of parameters to tune relative to block kriging to produce similar results, and their tuning process is fully metric-based and automated, capitalizing on the ML framework, which offers a workflow of inter-compatible and automatable methods (e.g., cross-validation and model selection). Other benefits include: (1) freely available and mostly open-source libraries that are tuned for high performance computation across a range of platforms; and (2) lower barrier of entry into spatial modeling for practitioners of artificial intelligence, ML and data science by reformulating block modeling into ML workflows. Hence, in addition to satisfactory performance characteristics, our method is more reproducible, less subjective and more accessible. Thereafter, we deploy our method to a simulated porphyry Cu deposit (Garrido et al., 2018, 2020) to create three dimensional (3D) geometallurgical block models.

SYNOPTIC REVIEW OF GEOSTATISTICS AND COMMON CHALLENGES

Geostatistics is a well-developed domain of geosciences with an extensive and impressive history, as well as a large variety of known applications (primarily in solid-earth science and related fields) that include mapping, resource modeling, spatial prediction, and interpolation and change of resolution (Krige, 1997; Ortiz & Emery, 2006; Talebi et al., 2019). Its fundamental premise is that many natu-

rally occurring spatial phenomena that are not subjected to a high driving rate of mixing and transport processes exhibit a spatial variability (essentially pseudo-equilibrium spatial distributions; Matheron, 1967). These spatial patterns can be quantified to produce models of the variability structure, which, in turn, becomes useful for estimation, mapping, general interpolation and other purposes (Isaaks & Srivastava, 1989; Isaaks, 2005). In other words, the statistics of sampled data of many solid-earth processes are expected to exhibit a general spatial variability, and therefore, correlatability. A particular task that is commonly performed using geostatistics is block modeling. Although block modeling is often used for resource estimation within the context of the mineral value chain, it is implicitly used for other purposes as well. This is because block modeling is a solution to the 'change-of-support problem' (Gelfand et al., 2001; Gotway & Young, 2002). The idea of a geostatistical support refers to the ideal dimensionality of a geospatial measurement, which could be 1D (a point sample), 2D (a surface response) and 3D (a volumetric response). Changing support refers to changing the representation of the geospatial phenomena from one type of support to another. Creating maps, for example, requires that a 1D support be changed into a 2D or even 3D support to represent an areal or volumetric distribution of some quantity. Change of support is an unsolved problem in the general sense, because there is no unique and universally applicable solution, and various disciplines such as geostatistics, produce a type of solution based on its capabilities and prior assumptions (Gelfand et al., 2001; Gotway & Young, 2002). The change-of-support problem is commonly encountered in geosciences due to two key characteristics: (1) native-support sampling, in which case data are collected at the exact spatial resolution at which the natural phenomenon occurs, is often hindered by sampling opportunity constraints and resource limitations; and (2) the necessity for scientific reduction and associated analysis techniques has been a significant driving factor in shaping data-engineering methods (Cressie, 1990; Carvalho and Deutsch, 2017). Although data-generation methods are currently evolving, primarily driven by an evolution in guiding scientific philosophy (from reductionism to system considerations) brought forth by the change in the purpose of geoscientific data (toward increasing use of transdisciplinary techniques, such as artificial intelligence, ML and data science).

A particular solution to the change-of-support problem for the purpose of block modeling is block kriging, although this concept is also relied upon across the broader geostatistics. Within the domain of geostatistics and for the purpose of resource estimation, this is the most commonly used method. Although block models are often visualized as quantized (discretized) representations of a physical phenomenon (e.g., a portion of an orebody), a range of discretization exists such that the perception of discretization is not always apparent (Vann et al., 2003). For example, in the use of geostatistical interpolation methods to create regional maps from geochemical concentrations, the effect of quantization can be either masked by large pixel sizes or post-hoc smoothing, and hence, block-level quantization is not always obvious (Abzalov & Humphreys, 2002). For resource estimation, quantization is usually visually obvious because absolute distances involved are typically much smaller than regional maps and discretized blocks favor extraction sequencing and reconciliation, for which smoothing is also undesirable (Sarma, 2009). Block kriging is therefore capable of creating 2D or 3D models of inferred reality at a range of quantization using typically 1D data. This process is followed under implicit or explicit assumptions: (1) that the 1D samples are infinitesimally small, such that their internal structure is omitted at the scale of observation (or effectively a 0-dimensional, punctual support); and (2) that samples are representative of spatial variability at the scale of observation. However, in practice, the correctness of the approximation of punctual support is a continuum and depends on the volume of the block versus the volume of the sample, and in the extreme case that the two quantities are comparable, punctual support assumptions are clearly violated (Matheron, 1967; David, 1976). While sample-points-to-block or block-to-block kriging is the industry-preferred method of mineral resource estimation, it is not without drawbacks. In many cases, poor implementation of kriging or subjective configuration of kriging parameters have led to the under- or over-valuation of regionalized variables (e.g., resources, Krige, 1997; Isaaks, 2005). One of the primary issues in point and block kriging is the number of (usually manual) steps involved, which can result in a time-consuming and challenging implementation process, even with some of the sub-tasks automated. Furthermore, point or block kriging often requires an expert-based tuning of various parameters, which is manual and can involve

subjectivity, in turn decreasing reproducibility. However, geostatistical modeling is not irreproducible. In practice, reproducibility is usually ensured via thorough rigorous training, documentation of the parameters and methodology. Moreover, advancements in geostatistical software and the establishment of guidelines for parameter calibration have contributed to standardizing and facilitating the parameter tuning process.

Despite substantial progress toward standardizing resource estimation workflows, there is still no universal protocol, and resource assessment remains a discipline in which only a few practitioners are considered "Competent and Qualified Persons[1]" due to the technical and mathematical difficulties involved. This is less of an issue for typical mapping uses of geostatistics because the knowledge required to estimate resources exceed merely spatial interpolation, although rigorous geostatistics is an entire discipline and is not typically within the training of modern data talents (e.g., data scientists). This is a solvable problem toward attracting data talents into geosciences, as we seek to demonstrate with our proposed method. The solution requires a geostatistical examination of ML algorithms. ML algorithms are used by data scientists but are incapable, in general, of a rigorous change-of-support in the manner of geostatistics (Veronesi & Schillaci, 2019). This is because ML algorithms (including many deep learning architectures) are unable to constrain local spatial structures. Hence, relationships between features and data labels generally do not capture changes in support, even if spatial coordinates were used as features. Indeed, for the activity of mineral prospectivity mapping, although many ML and deep learning algorithms have been used, all evidence layers (e.g., geochemical and geophysical data) are interpolated before modeling using non-ML approaches (e.g., Lawley et al., 2021, 2022; Parsa et al., 2023). Hence, predictions made using ML algorithms occur almost exclusively in the same support as features in data. This unfortunately means that in cases where input data points are considered punctual, but the output is intended for volumetric representation, ML is not expected to yield rigorous results and would generally not be able to change output resolution without retraining (e.g., relearning relationships at a different resolution). However, this view is over-simplified and ignores that the punctuality of support is an approximation necessary for geostatistical methods and that no physical sample can, in the strictest sense, become punctual. In effect, all physical samples are finite in volume and this fact can be exploited to assist ML algorithms to perform a rigorous change-of-support. The manner in which this could be performed is not unique, but a straightforward method is to also exploit the lack of support awareness in ML algorithms. The process consists of: (1) determination of an approximate support size; (2) making an at-scale prediction of a grid, which is sized such that each grid spacing is roughly the volume of the support (microblocks); and (3) down-sampling of microblocks to macroblocks to achieve the desired block size (integer down-sampling incurs zero re-sampling error). Our proposed method recognizes properties of geospatial data beyond the data-driven realm, because the representativity of the data in terms of sample volume and geostatistical considerations are clearly required and neither are strictly transdisciplinary knowledge by origin. This is also a general method, because any ML algorithm and geospatial data could be leveraged in this manner to directly produce block models at any desirable resolution. The punctuality assumption of geostatistics is transformed into a microblock assumption in our approach – that the size of microblocks must be sufficiently small compared to the size of macroblocks. In the case where the macroblock size is much larger than the support volume, it is not necessary that the microblocks are sized to roughly the support volume, because averages over an area of a large ensemble of predicted microblocks would still converge to the macroblock mean. In this manner, the approximation of punctuality carries over into the ML-based method, but the implications are different to those of geostatistics. Our approach is independently conceived but is mechanistically similar to some implementations of block kriging, which eschews the ML framework for geostatistical algorithms and workflow designs.

DATA AND METHODS

Synthetic Data Generation

The synthetic data were generated using the sequential Gaussian simulation function in GSLIB (Deutsch & Journel, 1992), via the GeoStatsPy

[1] A Competent Person is a minerals industry professional (i.e., registered as a professional of appropriate membership class and organisation including recognised professional organisations) with enforceable disciplinary processes including the powers to suspend or expel a member. Rules and regulations vary across countries.

interface in Python (Pyrcz et al., 2021). The generation process consisted of: (1) creating a parameter grid that included the sample rate, nugget effect and anisotropy; (2) holding all parameters but one as experimental control and varying the single parameter as the experimental variable to produce a variogram model; (3) generating a realization using sequential Gaussian simulation; and (4) sampling the realization at a specific grid-sampling interval. This procedure allowed us to perform a controlled exploration of the parameter space by isolating the effects of each parameter on our proposed approach versus that of block kriging within the parameter space. The anisotropy parameter was defined as the ratio of the minor and major ranges, such that if the anisotropy parameter were set to 1, the major and minor ranges would be equal. In this context, 'sample rate' refers to the density or frequency of data points or samples collected in the spatial domain. It represents how closely spaced the data points are within the area being sampled. A higher sample rate corresponds to more closely spaced data points, whereas a lower sample rate indicates more widely spaced data points. Two types of sampling methods were explored – regular sampling, which used a fixed spacing to designate sample sites, and biased random sampling, which used a smaller combination of a fixed spacing and an additional number of sample sites that are randomly scattered. The nugget effect was explored by varying the nugget parameter. Modulating the nugget effect of the synthetically generated data allowed us to capture physically realistic microscale variability and/or measurement error, which permitted us to understand the capability of our proposed method with increasing nugget effect. The entire parameter grid is given in Table 1. The other key fixed parameters included minimum and maximum grid values (− 3.0, 3.0); grid size (100 cells by 100 cells); grid spacing (10 cells); sill (1.0); azimuth (135°); major range (800 cells); and variogram model (spherical). As the grid exploration was performed via single parameter sweeps, the default (fixed) nugget effect was 0.1 and the anisotropy was 0.625 during the sample rate sweeps. During the nugget effect sweep, the fixed sample rate was a spacing of 30 cells and the anisotropy was 0.625. Lastly, during the anisotropy sweep, the fixed sample rate was a spacing of 30 cells and the nugget effect was 0.1. This was repeated for biased random sampling method with the sample spacing replaced by its corresponding value (see Table 1) for the number of random samples.

Machine Learning-Based Block Modeling

Formulating the block modeling problem into a ML task requires two key considerations. First, predictions must be made at data label-volumes that are substantially smaller than the desired block size, such that they approximate the condition of a static support. This satisfies the constraint that almost all ML algorithms are unable to perform dynamic changes-of-support. Second, spatial coordinates are to be used as the sole features for spatial estimation tasks. No model deployment would occur beyond sampled areas (hence no spatially transferred learning) because the models would be used solely for interpolation. To satisfy the first condition, the desired block size must be re-discretized into smaller blocks, which we call 'microblocks'. After predicting all microblocks, they are averaged (via spatial down-sampling) to produce a macroblock model, which is specified at the desired block size to perform a change-of-support. This type of down-sampling is known as a type of spatial signal compression (Crochiere and Rabiner, 1983). Additional knowledge of signal processing is useful in our approach but is not a critical component. The key signal processing consideration is that the down-sampling should ideally be of an integer factor to incur no re-gridding error.

For microblock grid spacing, we adopted a linear spacing of 1/10 of that of the block model produced by the synthetic data generation process (e.g., each macroblock comprises 100 microblocks). This is a heuristic setting in our case to satisfy the approximation that the size of microblocks is much smaller than that of macroblocks, because there is no effective volume associated with synthetic data. Based on our synthetic data, there was no evidence of increasing accuracy of macroblocks with finer microblocks. In general, documentation of actual data generation processes would indicate the sample volume (e.g., Armstrong & Champigny, 1989; Cressie, 1990; Annels, 1991). The microblocks were then predicted using a ML algorithm and subsequently, macroblocks were produced by down-sampling the microblocks by an integer factor of 10, retrieving the original block configuration of the synthetic data (100 cells by 100 cells grid size with 10 cell blocks).

In a strict technical sense, ML algorithms, as a whole, are not generally intended for spatial modeling, because they were originally designed for feature space characteristics that encompass essen-

Table 1. Parameter ranges for the explored parameters and their interval construction methods

Parameter	Parameter range	Interval construction
Sample rate – regular sampling	1 per 23 cells to 40 cells	1 cell intervals
Sample rate – biased random sampling	625 to 1736 cells total	Area of grid divided by the square of the regular sampling rate
Nugget	0 to 0.70	Intervals of 0.05 (out of a total sill of 1.00)
Anisotropy	0.1 to 1.0	20 evenly divided intervals

tially all types of data. Of the entire superset of data characteristics, very few are typical of spatial data, which makes ML, in general, unspecific to spatial learning tasks. For spatial data, some common characteristics include: (1) generally low feature space dimensionality (usually 2 to 3 dimensions in non-temporal data); (2) explicit connotations of spatial feature interactions, e.g., in the form of spatial (linear) correlation; (3) primarily linearly independent feature coordinates (e.g., the Cartesian coordinate system and the use of the Euclidean metric to measure distance); (4) low sample rate (resulting in sparse samples) in the case of manual data generation; and (5) potentially noisy data with a spatial noise distribution (e.g., high nugget effect). In particular, although characteristics (1), (2), (3) and (4) do not preclude a variety of ML algorithms, they ameliorate the advantage of many algorithms by design. For example, by design and usage, tree-based methods and artificial neural networks are intended to better model data with high-dimensional, complex, and nonlinear feature interactions. If they are used with trivial data, such as those that would warrant a linear regression, then their algorithmic complexity would not translate into realizable performance benefits. In addition, the increase in model parameters would render the models and results difficult to interpret, and extrapolations likely unreliable. The desire to extrapolate is also a stronger consideration in geostatistics than in ML, especially since spatial correlation permits some extent of knowledge outside of sampled areas and many ML algorithms extrapolate poorly or not at all (e.g., tree-based methods). These characteristics imply that a relatively small variety of known and common ML algorithms may be useful for block modeling without algorithmic modifications.

Based on intuition from geostatistics, there are two classes of common ML algorithms that are theoretically appropriate for this task: neighbor-based (e.g., k-nearest neighbors or kNN; see Fix & Hodges, 1951; Cover & Hart, 1967; Witten & Frank, 2005; Kotsiantis et al., 2007) and Gaussian processes (Rasmussen & Williams, 2006; Kotsiantis et al., 2007). This is because neighbor-based methods use averages of neighbors, which is similar to inverse distance modeling, and the Gaussian process is a ML generalization of kriging, although without as many hyperparameters, and hence, somewhat less flexible. Other methods include a variety of tree-based methods, such as random forest and boosted methods (e.g., Ho, 1995; Breiman, 1996a, b; Freund & Schapire, 1997; Kotsiantis, 2014; Sagi & Rokach, 2018). These methods are not suitable for block modeling in the traditional sense, because they do not include geometric characteristics in the feature space, and in the presence of sparse samples (which is the case of most geoscientific sampling), they create orthogonal decision boundaries in lower number of dimensions and, therefore, result in cross-hatched patterns in block models that are unrealistic for our purpose, at least without post-processing (e.g., Zhang et al., 2023). Additionally, tree-based methods excel at leveraging nonlinear feature interactions in high-dimensional feature space, which is a benefit that would not be easily realizable for spatial modeling in 2 or 3 dimensions and in cases in which the variability is dominantly linear (e.g., as modeled through a variogram or correlation matrix and because sampling through volumetric averaging over space introduces linearity through the central limit theorem, see e.g., Hsieh, 2002). The kNN algorithm uses a variably weighted and averaged value of the nearest "k" neighbors (a model hyperparameter) to determine the value of an unknown data point. This is very different to inverse distance modeling because distance is not a fixed parameter for kNN. In this paper, we focus on the systematic evaluation of the kNN algorithm for block modeling as a proof-of-concept. The hyperparameter grid for the kNN algorithm was k = (2 to 30 in intervals of 1) and the weighting method was based on the inverse neighbor distance. During model selection, 20% of the sampled data was re-

served for testing and the remainder was used for model training. During hyperparameter tuning, 4-fold cross-validation combined with the coefficient of determination (CoD or R^2) metric was used to select the best model. The implementation was in Python using the Scikit-Learn library (Buitinck et al., 2013).

Geostatistical Block Modeling

Geostatistical block modeling is a very mature task in spatial data modeling. Here, we provide a simplified outline of the process for brevity. To perform geostatistical block modeling, we used a standard geostatistical spatial estimation approach based on ordinary kriging (OK), which was implemented in GSLIB (Deutsch & Journel, 1992) and Gestates (Pyrcz et al., 2021). OK is a method of spatial interpolation that uses a weighted linear combination of sample values at nearby locations to estimate the value at an unsampled location. The kriging process begins by defining a set of sample locations, known in geostatistics as the 'neighborhood' or 'search neighborhood', and the corresponding values of the variable at those locations (Krige, 1951). It is worth noting that the vernacular from various geoscientific disciplines overlap and sometimes conflict with those of artificial intelligence, ML and data science. Neighborhood in the sense of the ML algorithms does not refer to a spatially constrained neighborhood, in general, and its context depends on the definitions of the features. The kriging estimate at an unsampled location is then calculated as a weighted average of the values at the sample locations within the neighborhood (Goovaerts, 1997). The weights are calculated such that they minimize the estimation error, which is measured by the variance of the residuals (Krige & Magri, 1982). In the case of OK, the sum of the weights for the individual samples is constrained to unity and a Lagrange multiplier method is used to find a solution. Thus, the known mean of samples is not required in OK. There are also underlying assumptions, such as the stationarity of domains and some extent of subjectivity in the form of modeling choices. These can greatly impact the results and the interpretation of the kriging estimates.

There are a number of standard metrics in kriging to profile resulting model performance. These include: (1) kriging variance, which is a measurement of the precision of an estimate at a given location (Krige, 1997); (2) kriging efficiency (KE; Krige 1997; Deutsch & Deutsch, 2012), which is a metric of the effectiveness of kriging estimates; (3) kriging slope of regression (e.g., Deutsch et al., 2014), which is a metric used to estimate the local slope of a variable of interest in spatial interpolation. Based on a number of case studies, there is a correlation between the KE and the kriging slope of regression (Krige, 1997). However, these metrics are strictly within the discipline of geostatistics and no direct counterpart of them exist in ML. Hence, they cannot be used to cross-compare ML- and OK-based models. In addition, these metrics also cannot be used to compare models with the synthetic ground truth. For this reason, we did not employ standard kriging metrics in this study. Instead, we made use of metrics from ML to compare models against ground truth.

In this study, we employed block OK with block parameters that matched with the synthetic data generation, and also, notably, the use of a spherical model (Krige, 1997; Olea, 1999). This helped to maximize the performance of block OK in a synthetic study setting, because the data were synthetically generated using a spherical variogram model; choosing a spherical variogram model during block modeling maximizes the accuracy of the resulting block models. However, in general, no ground truth is available and the choice of the variogram model is heuristic. The variogram model is a key subjective choice in a typical geostatistical workflow for interpolation because its structure cannot generally be deduced unambiguously from data alone and no general knowledge is universally reliable. Maximizing the performance of block OK is important to fully appreciate performance contrasts between the traditional OK approach and our proposed ML-based approach. The tuning of model hyperparameters is automated and includes optimization of the number of neighbors to use and fitting the variogram following Zhang et al. (2023). In particular, the number of neighbors was selected using the elbow method combined with the mean absolute error metric, and the variogram was fitted using an automated least-squares method with a linear bias toward shorter distances to favor better fitting of the pre-range portion of the spherical model.

Performance Assessment

To assess the performance of both ML-derived and geostatistical block models by comparing the

resulting block models against the synthetic ground truth, we used a variety of metrics that included the coefficient of determination (CoD or R^2), the mean absolute percentage error (MAPE, although it is expressed as a fraction by default, which was the case in this study as well; see https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_percentage_error.html), and the median absolute error (MedAE; see https://scikit-learn.org/stable/modules/generated/sklearn.metrics.median_absolute_error.html). In addition, we examined key statistical moments that included the mean and standard deviation (STD) of all models. Lastly, to assess the degree of smoothing in block models, we used a linear version of the dynamic range metric, which is defined as the difference between the highest and the lowest block values. The dynamic range metric is usually used in sensor engineering, acoustic and image processing to measure sensor and digital processing capability (https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=723-03-11). In our case, we can treat the resulting block models as images. Hence, the dynamic range metric captured the maximum contrast of block models, which is a useful metric of the level of smoothing because smoothing reduces intra-model contrast. To ensure that the results were robust, we performed 50 runs for each combination of parameters and averaged the results, with 50 synthetic datasets being generated by varying the random number generator's seed. During each run, a single workflow was used, which created reproducible synthetic data conditions, including the random number seed and sampling pattern. The samples were then used to perform block modeling and performance assessment. Metric results were thereafter averaged over the 50 runs to produce a statistically robust outcome.

Method Deployment Case and Data Description

Geometallurgy is the integration of geological, mineralogical, financial and metallurgical data (especially extractive metallurgy) in (usually) 3D space to create a spatially-aware predictive model of mineral processing. The benefits of geometallurgy in mining engineering are significant, including improvements in ore quality, mine planning, plant performance, cost reduction and product quality improvement (Jackson et al., 2011; Ortiz et al., 2015; Dominy et al., 2018; Garrido et al., 2019). To incorporate these benefits into the mining value chain, metallurgical responses and proxy variables need to be included in the block model, which is essential for optimizing mine planning and downstream mineral processing. This enriched block model, known as the 'geometallurgical block model' (GMBM), is based on the transfer of simulated attributes (Deutsch et al., 2016; Garrido et al., 2017, 2020). As there is now a general awareness of the lengthy exploration-to-extraction timelines and that a majority of the energy cost lies in beneficiation for many commodities and deposit types, there is a corresponding desire to drastically integrate the mineral value chain and enhance its agility. An early availability of GMBM is likely to be key to mineral value chain integration and enhanced agility because it permits downstream metallurgical process designs early and, therefore, design of industrial processes and construction of facilities could be implemented early. However, access to large mining exploration and geometallurgical databases for method development and academic purposes is difficult due to confidentiality restrictions or budget limitations. To address this issue and for the purpose of an academic study, we used a synthetic geometallurgical database of a typical porphyry Cu deposit (Garrido et al., 2020) for a prototype deployment of our method. This database contains geological and spatial coordinates, mineralogical data, chemical assays and geometallurgical responses, such as rougher recovery and Bond work index (BWI). The BWI is a measure of the resistance of the ore to grinding using a laboratory ball mill. It is a widely accepted measure of ore grindability, indicating the energy required to grind the ore to a specified size. The BWI is expressed in units of kilowatt-hours per ton of crushed material (kWh/t_c). The mineralogical data include the proportions of chalcopyrite, bornite, grouped clay minerals and pyrite in volume percentage (vol.%), whereas chemical data include Cu grade in weight percent (wt.%). Boreholes at various angles and lengths were extracted from the simulated deposit and processed for further modeling, such as the spatial distribution of minerals, elements, and geometallurgical responses. Figure 1 shows the configuration of boreholes used for subsequent modeling. It is important to note that many geometallurgical variables are non-additive, such as comminution indices, and combinations of compositional and non-compositional data (Deutsch et al., 2016). Directly interpolating non-additive variables can result in misleading interpretations, although non-additive variables may not have the same im-

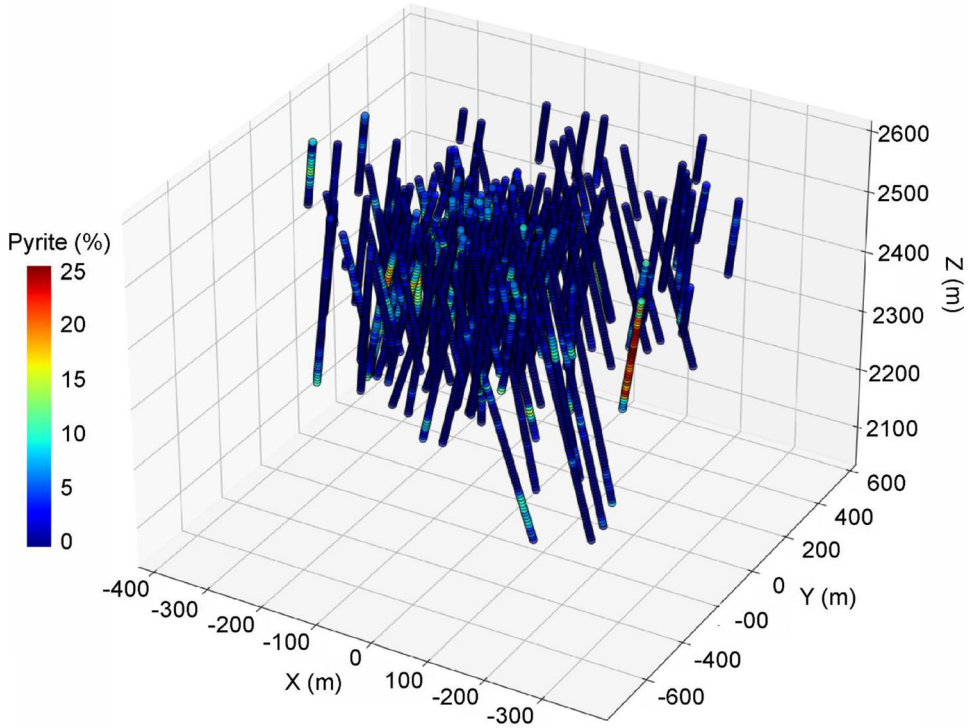

Figure 1. Graphical visualization of oriented boreholes with pyrite (vol.%) depicted (n = 153) with variable azimuth and dip.

pact on all ML algorithms. However, for the deployment portion of this study, the focus was not to validate our method for non-additive variables in general, but to provide a quantitative comparison and examine the feasibility of our method in 3D using physically realistic data.

RESULTS

Regular Sampling

Qualitatively, under regular sampling conditions, block OK produces models that are generally higher in contrast (Figs. 2, 3, 4, 5). However, there is a visually discernible loss of fine-scale detail (e.g., Fig. 3) in the kriged models as compared with that of kNN models. The difference is noticeable in Figure 2 (compare the hotspot in the lower left corners). However, it was not obvious whether the increased detail retrieval resulted in overall more accurate models, as it may be accompanied by a proportionate increase in noise as well. An observable qualitative difference was that block OK models were visibly smoother in the interiors of high and low concentration regions, as compared with the ground truth and the macroblocks (e.g., blue and red zones in Figure 2). The qualitative differences became larger between block kriging and kNN results at higher nugget effects (e.g., Fig. 4). Increasing levels of anisotropy did not appear to exhibit a systematic qualitative effect (e.g., Fig. 5).

Quantitatively, with increasing sample spacing, there was an appreciable performance loss in both kNN and kriged models, although the CoD, MAPE and MedAE metrics indicated that, on average, the kNN models were closer to the synthetic data (Fig. 6). The dynamic range was systematically higher in the kriged models, which was consistent with the departure of STD scores from those of the samples (Fig. 6). This observation was consistent with the qualitative finding that kNN models exhibit less contrast compared to kriged models. In general, with increasing nugget effect, there is a gradual loss in performance of both kriged and kNN models, although the losses in CoD and MedAE are more extensive with the kNN models (Fig. 7). The dynamic range and STD scores also degraded slightly

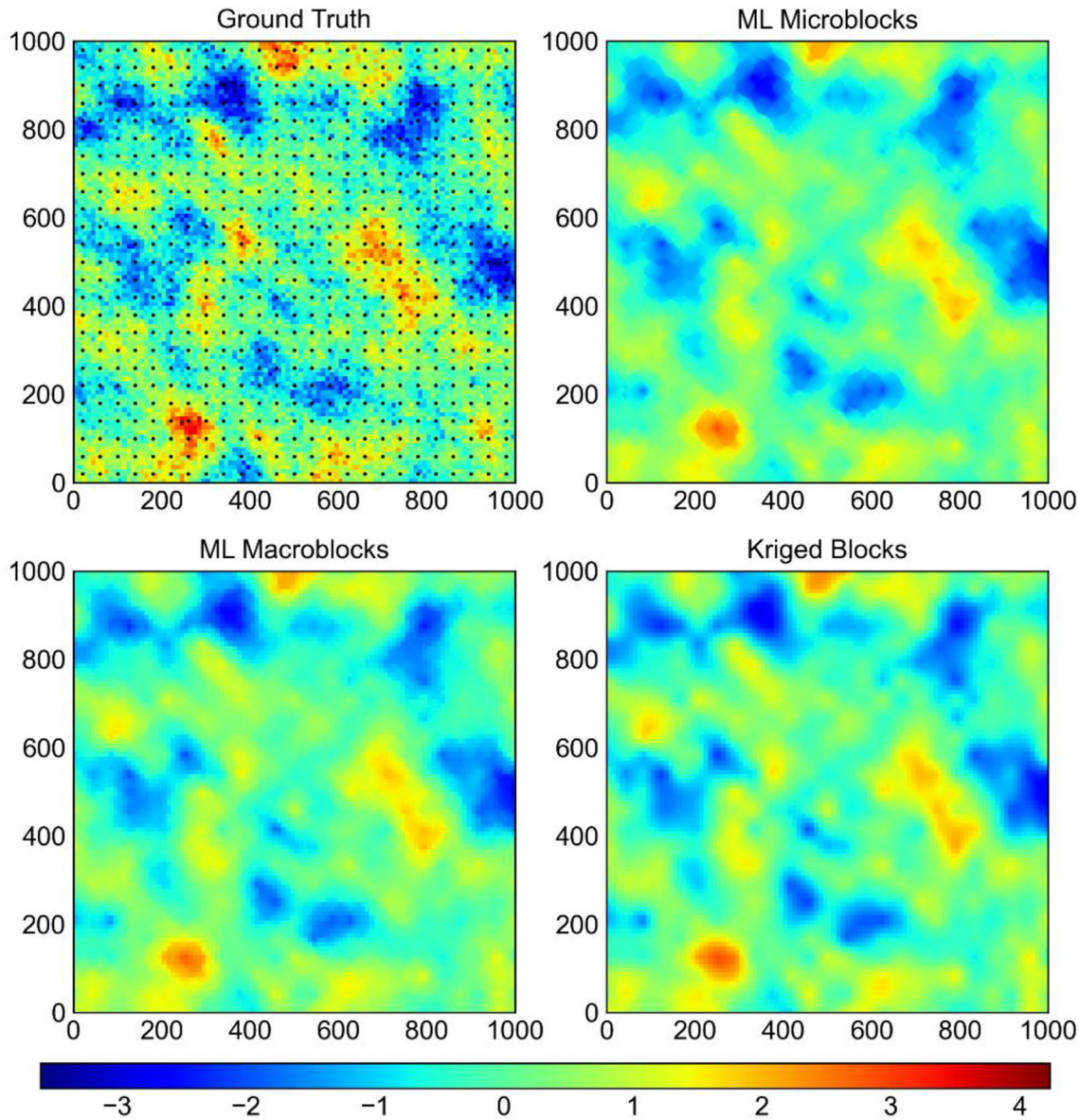

Figure 2. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under regular sampling conditions. Spacing between samples is 40 cells.

more rapidly with the kNN models as compared to the kriged models (Fig. 7). There seemed to be no robust systematic patterns of either the kNN or kriged models in the metric scores with changes in anisotropy (Fig. 8). These findings indicate that under regular sampling conditions, block OK yields generally more accurate models but at the expense of a loss of fine detail. For other applications, it would be possible to tune the kriging neighborhood to emphasize local detail, although this is not generally automatable. In contrast, the kNN-based block modeling method was less accurate at the distribution mean and median but it was capable of retrieving more fine-scale detail, although it seemed to suffer from more smoothing despite increased levels of fine detail.

Biased Random Sampling

The results of the kNN and kriged modeling under biased random sampling conditions were lar-

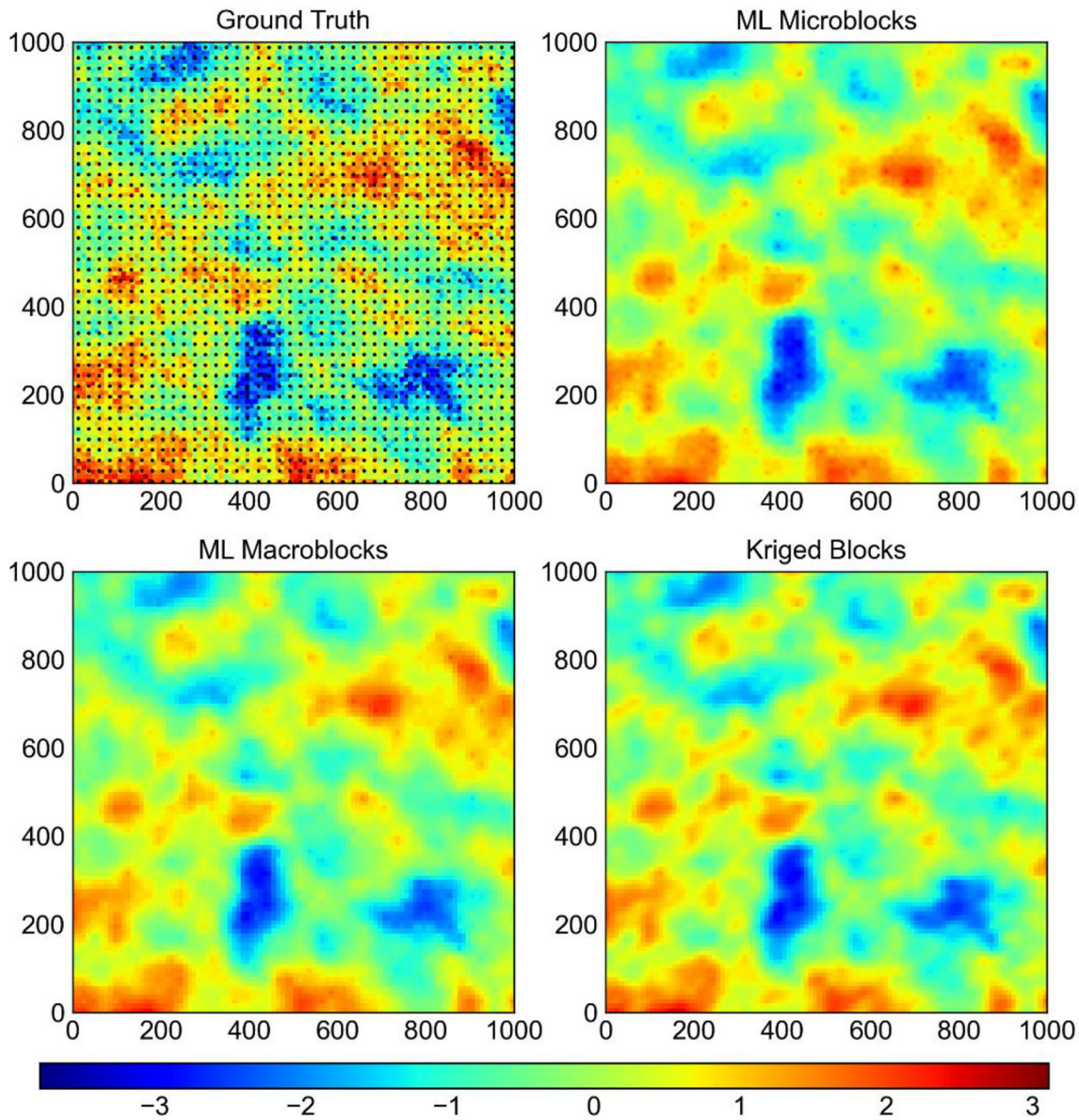

Figure 3. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under regular sampling conditions. Spacing between samples is 24 cells.

gely similar to those of regular sampling (Figs. 9, 10, 11, 12, 13 compared with Figs. 2, 3, 4, 5). In some cases, with strong anisotropy, the kNN models visually outperformed the kriged models qualitatively in the sense that they reproduced variability better across all azimuths (e.g., Fig. 13 but also less pronounced in Fig. 12). This effect was not as obvious under regular sampling conditions. Quantitative results were interesting in the sense that they were the reverse of those under regular sampling conditions. In particular, with changes in the number of random samples, the kNN models were able to retrieve a greater dynamic range and more accurate STD values, at the expense of lower CoD, MAPE and MedAE scores (Fig. 14). This was generally true as well for the nugget effect sweep, although at very low nugget values (< 0.1), the dynamic range performance of the kNN and kriged models were reversed in ranking but similar (Fig. 15). Although qualitatively obvious differences were noticeable

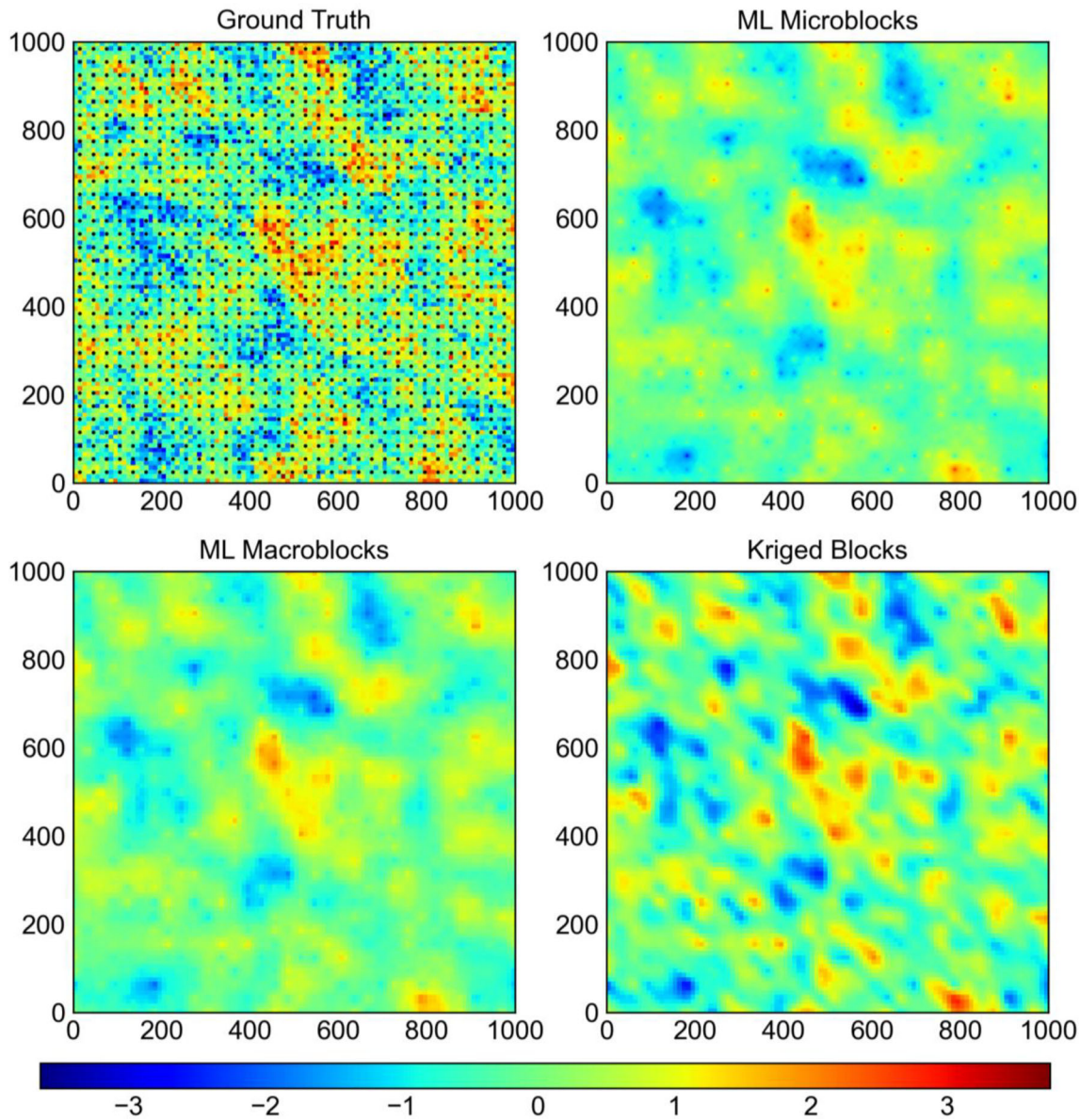

Figure 4. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under regular sampling conditions. Here, the nugget effect corresponds to 0.65.

with high anisotropy (e.g., Fig. 13), the metric scores did not indicate a substantial or measurable difference (Fig. 16). These findings indicate that under biased random sampling conditions, the kNN-based block modeling method is able to generally retrieve more detail but at the expense of also more noise. However, unlike the quantitative results under the condition of regular sampling, the dynamic range and STD scores of the kNN models were systematically better than those of the kriged models.

Method Deployment

The synthetic data used in this study had a clear distinction between the different ore minerals, such as bornite and chalcopyrite, enabling block modeling of each mineral volume proportion. The proposed microblocking approach demonstrated its capability to reproduce the data composition at the sampled and block location during the method development stage. During the visualization of the

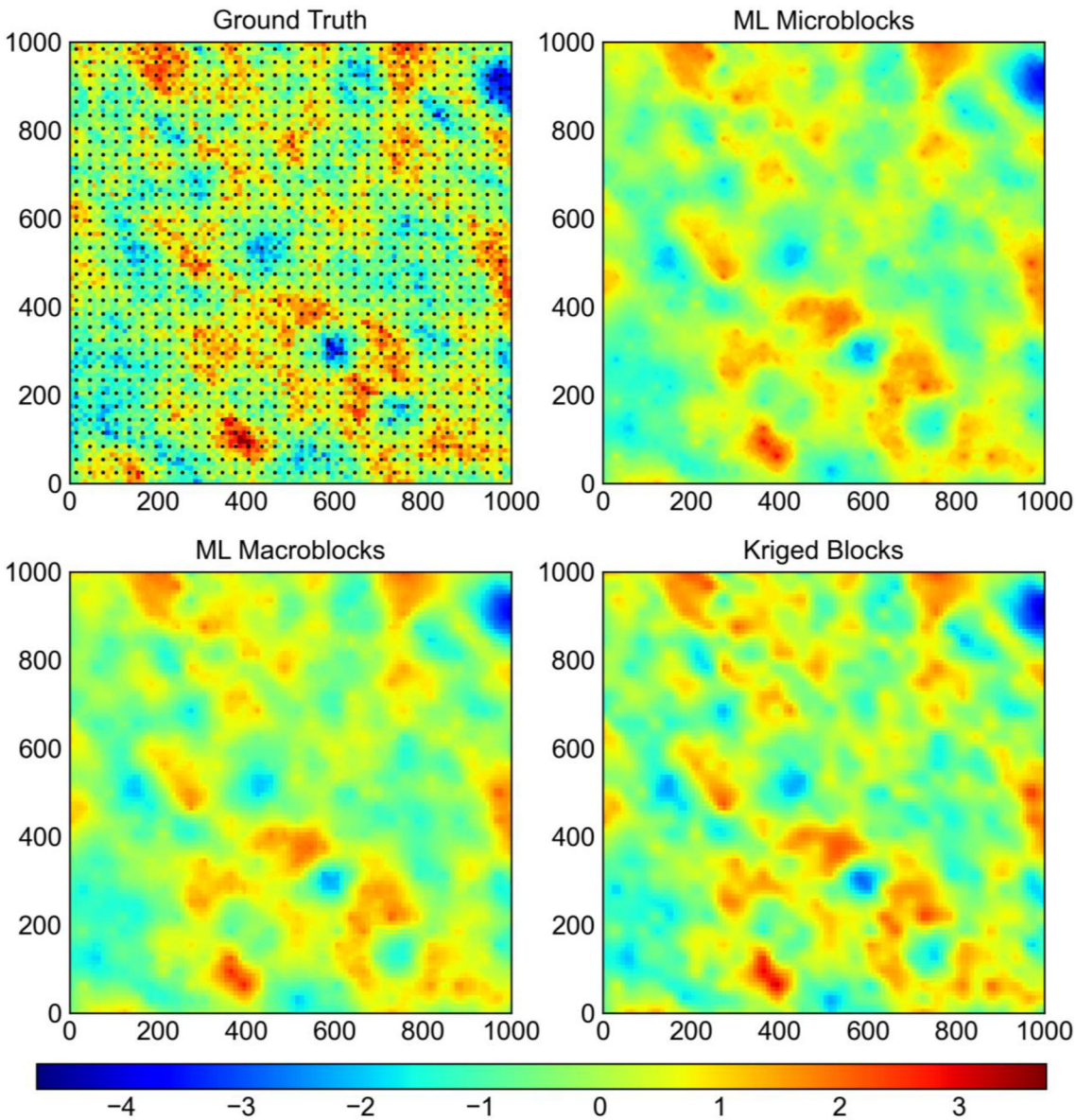

Figure 5. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under regular sampling conditions. Anisotropy corresponds to 0.1.

3D models, a 'moiré pattern' or 'moiré fringes', which are large-scale interference patterns that can occur when a partially opaque ruled pattern with transparent gaps is overlaid on another similar pattern, were observed. This was not related to the modeling methods prior to visualization and it was strictly related to visualization of repeated or stacking patterns of 3D models when viewed in 2D with transparency. To mitigate this effect, we employed a scale-factor-based visualization technique that allowed higher material concentration to have larger cubes and lower concentration material to have smaller and more transparent cubes. Figure 17 illustrates the predicted synthetic chalcopyrite proportion (vol.%) and its distribution in the studied porphyry Cu deposit. The results of the prediction indicate that chalcopyrite is heavily concentrated toward the center of the deposit and on the surficial frontier of the deposit. The average chalcopyrite content of the samples was 0.736 vol.% and it was

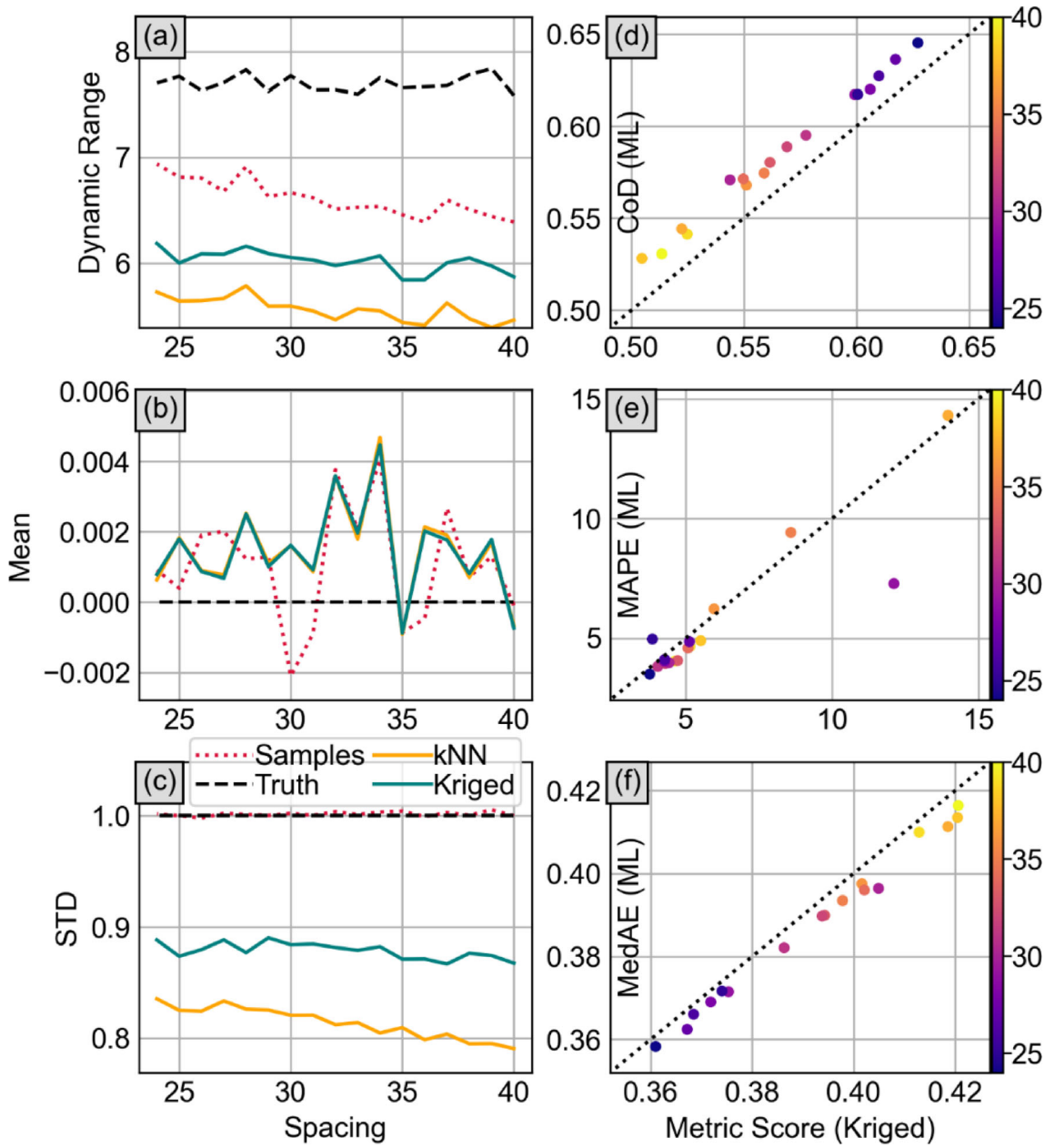

Figure 6. Parameter sweep of spacing (in cells) of regular sampling and impacts on (**a**) dynamic range, (**b**) CoD, (**c**) mean, (**d**), MAPE, (**e**) standard deviation, and (**f**) MedAE.

comparable to the ML estimate of 0.739 vol.% (Table 2). These findings are consistent with the simulated boreholes and ore deposit. The faint colors at the margins of the exploration area clearly illustrate the extent of mineralization (Fig. 17).

The microblock modeling approach employed in this study also allowed us to investigate the distribution and concentration of bornite in the porphyry Cu deposit. Our analysis revealed that bornite exhibits similar spatial patterns to chalcopyrite, albeit with different grade distribution regularities (Fig. 18). The average bornite content of sampled boreholes was 0.124 vol.% and the value was reproduced to 0.124 vol.% in the ML estimates, with

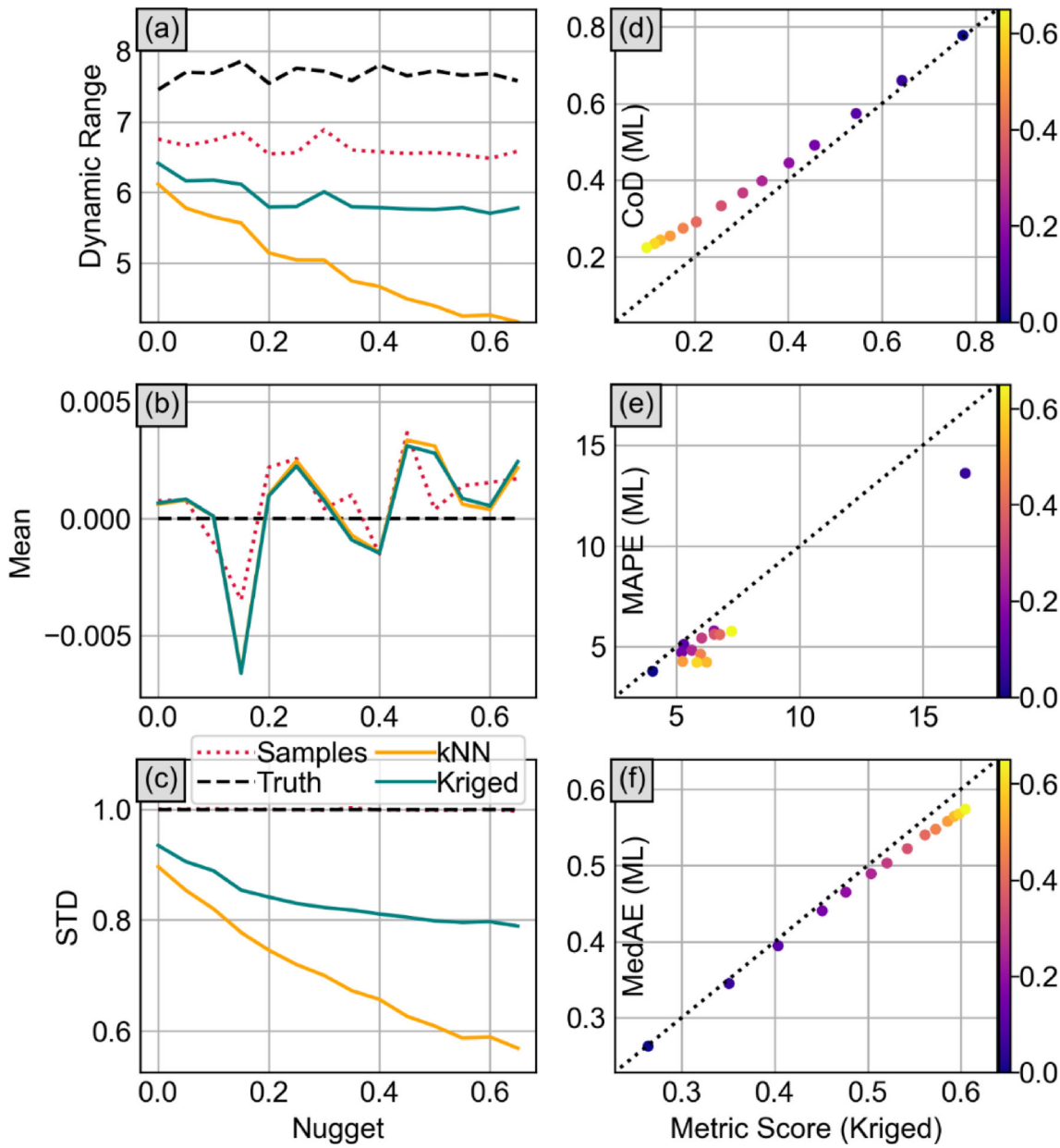

Figure 7. Parameter sweep of nugget effect of regular sampling and impacts on (**a**) dynamic range, (**b**) CoD, (**c**) mean, (**d**), MAPE, (**e**) standard deviation, and (**f**) MedAE.

the highest content clusters co-located in areas where the content of chalcopyrite is high (Table 2). These findings are consistent with the known distribution of the ore minerals governed by changing physico-chemical conditions during mineralization and can be explained by a sequential paragenetic sequence. In porphyry Cu deposits, chalcopyrite and bornite form by magmatic-hydrothermal processes (Tosdal and Richards, 2001; Richards, 2003). These processes involve a phase separation in the fluid following a pressure-release caused by hydraulic fracturing of the country rock. During this process, the Cu-rich fluids are generated by fractional crystallization of the magma, and subsequent migration through fractures and faults in the surrounding rocks deposit chalcopyrite and bornite. The formation of

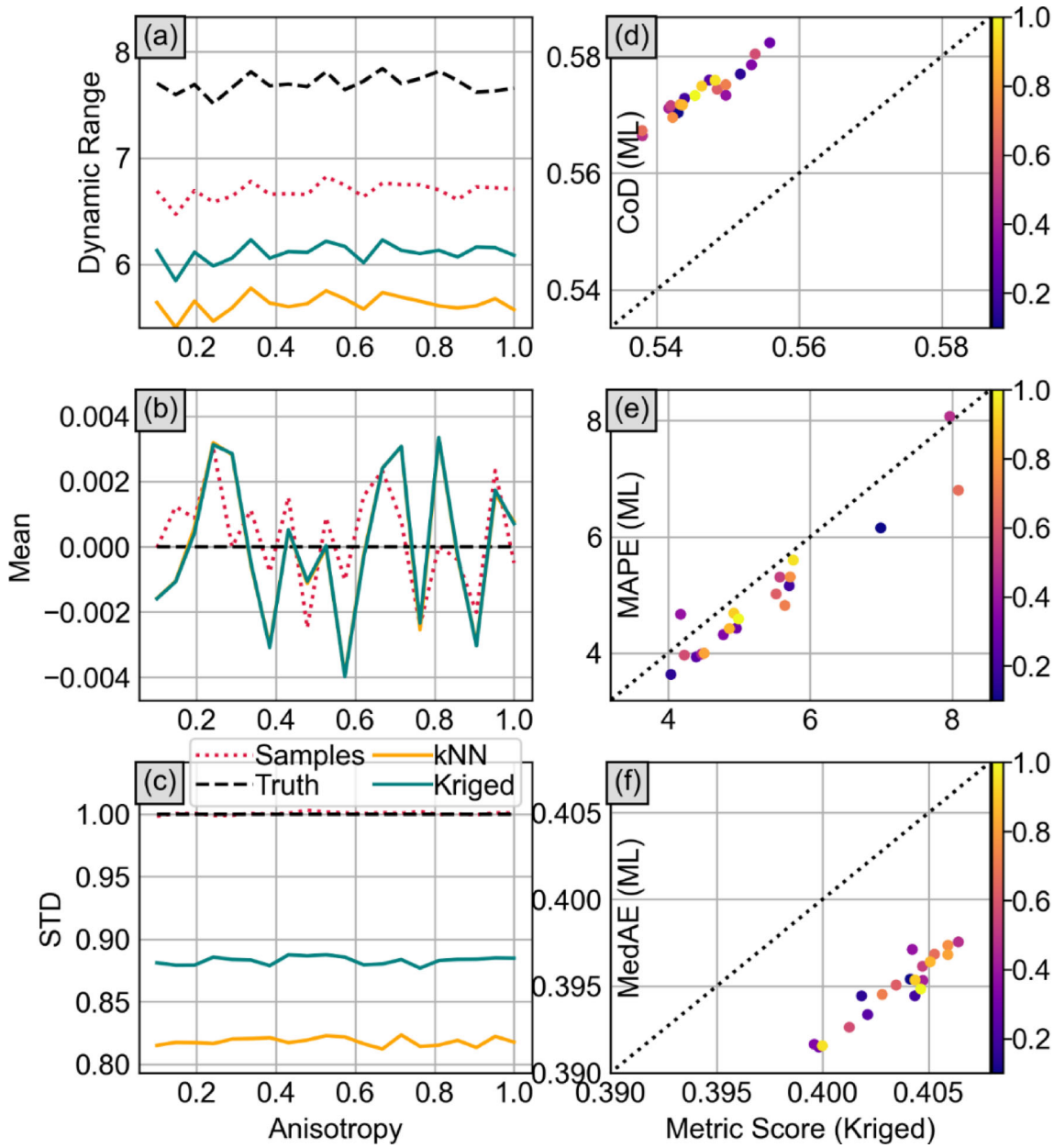

Figure 8. Parameter sweep of anisotropy of regular sampling and impacts on (**a**) dynamic range, (**b**) CoD, (**c**) mean, (**d**), MAPE, (**e**) standard deviation, and (**f**) MedAE.

chalcopyrite typically occurs earlier in the paragenetic sequence, followed by the formation of bornite. As such, bornite content in a given area can be linked to chalcopyrite content in that same area (Fig. 18).

The Cu grade in a porphyry Cu deposit is strongly related to the presence and concentration of Cu-bearing minerals. In the simulated porphyry Cu deposit, 3D modeling of the Cu grade (wt.%) showed that the host rock is enriched in Cu. The average Cu content in the sampled boreholes was 0.369 wt.% and it was comparable to the predicted estimate of 0.371 wt.% (Table 2). The distribution of Cu appears to follow a typical porphyry-style distribution pattern (Fig. 19). Porphyry Cu deposits typically contain Cu minerals such as chalcocite,

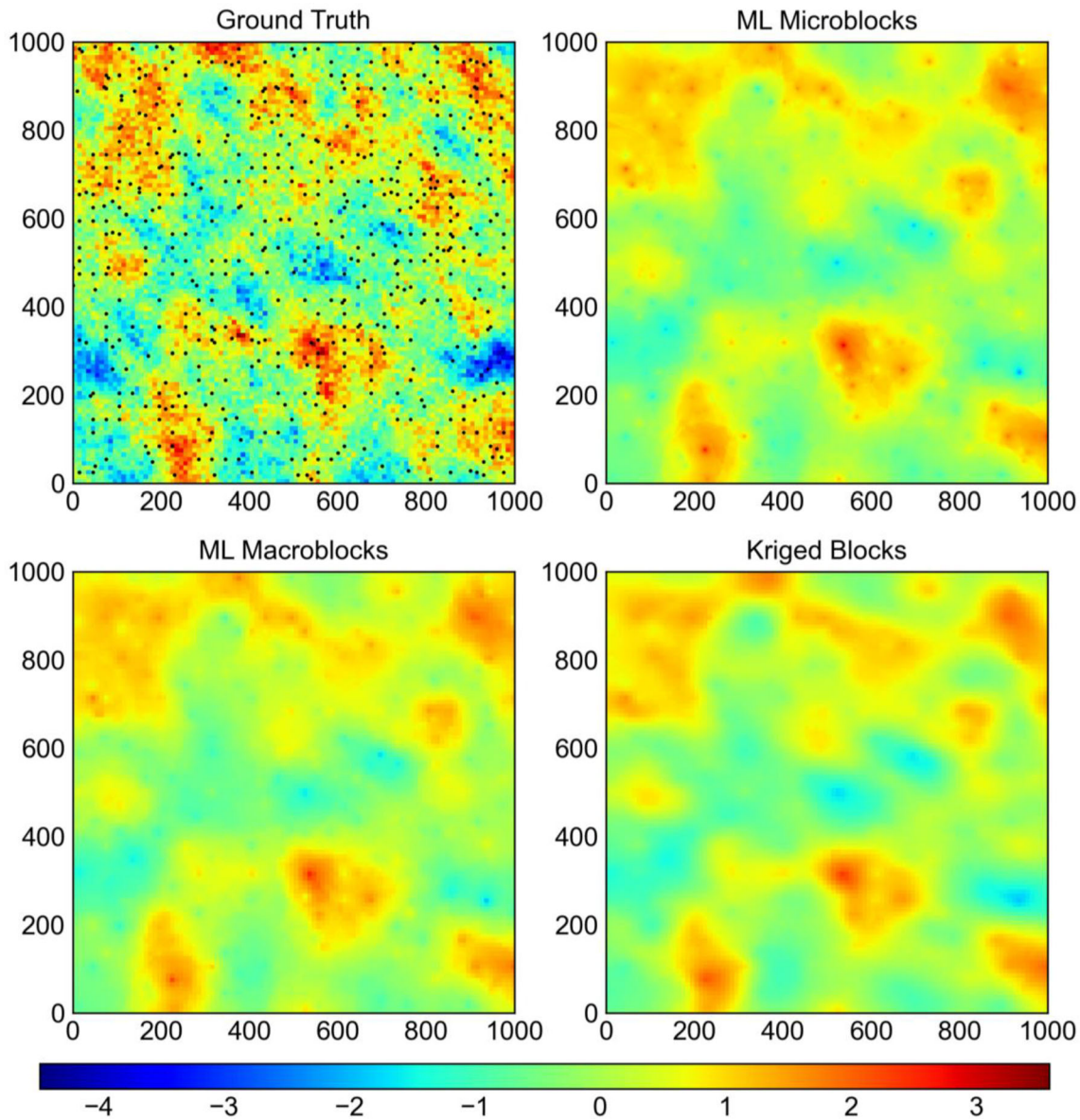

Figure 9. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under biased random sampling conditions. The number of samples equals 625.

bornite, chalcopyrite and tennantite. Among these minerals, chalcopyrite and bornite are considered the primary Cu-bearing minerals, and their concentration is critical for Cu recovery during metallurgical processing. These minerals typically occur as disseminations or veinlets within the host rock. Understanding Cu deportment to different ore minerals is essential in processing of porphyry Cu deposits. Bornite is considered a significant contributor to Cu recovery due to its high Cu grade and amenability to separation using froth flotation. The deportment of Cu to different ore minerals is an important consideration in metallurgical processing. For example, Cu minerals such as chalcopyrite are often associated with pyrite, which can create problems during processing due to the need for selective flotation. Chalcopyrite is also more difficult to recover due to its low natural floatability, high reactivity to oxidizing agents, and association with pyrite and other gangue minerals. Additionally,

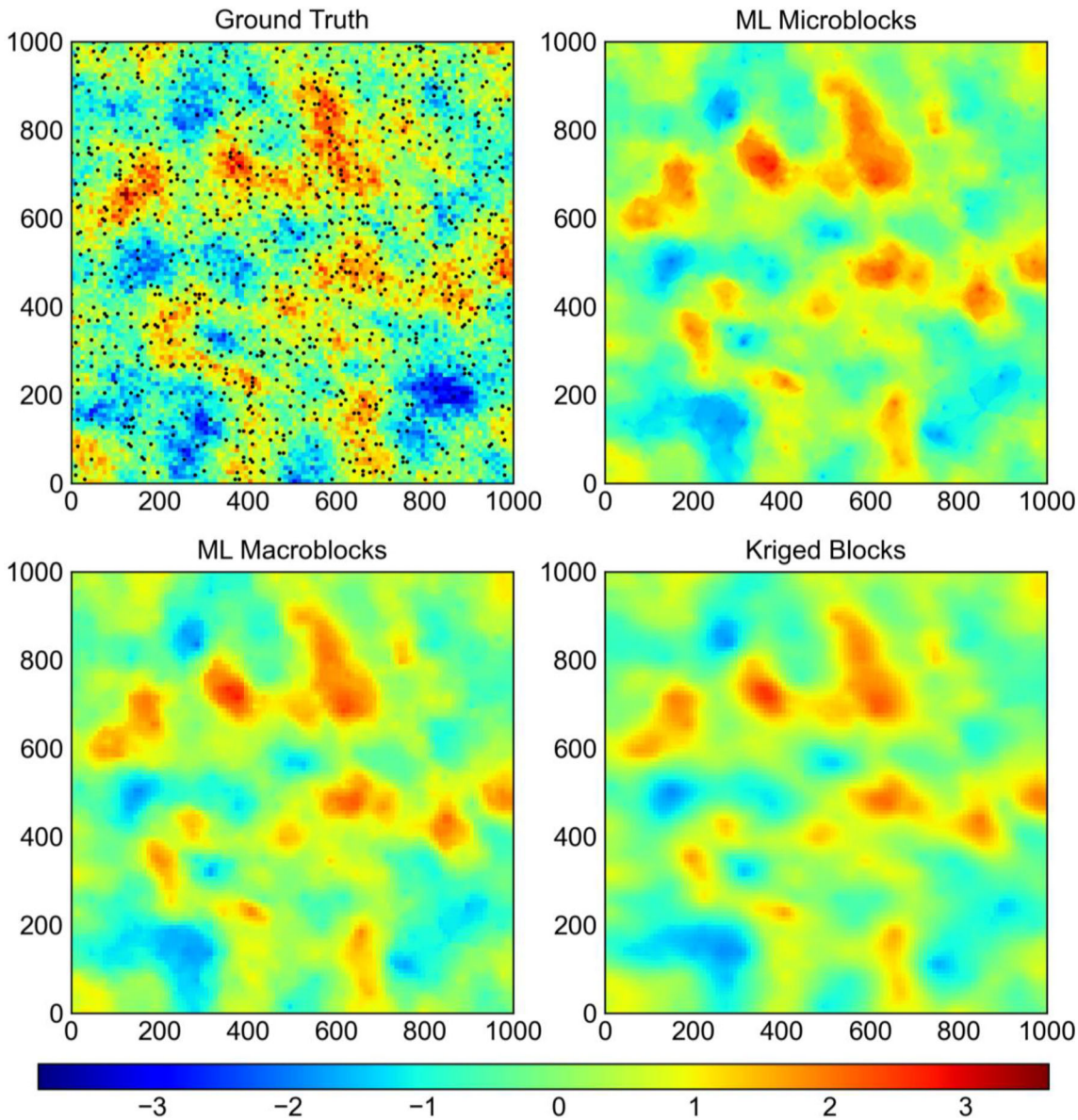

Figure 10. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under biased random sampling conditions. Number of samples equals 1736.

other minerals such as molybdenite can also impact the metallurgical behavior of Cu in the ore. Therefore, understanding the distribution of minerals and deportment of Cu is essential for optimizing metallurgical processing and maximizing Cu recovery.

Clay group minerals are known to adversely affect mineral processing, particularly flotation, due to their ability to adsorb reagents and reduce mineral recovery. In the simulated porphyry Cu deposit, clay minerals are present, accounting for up to 15 vol.% of the total mineralogy of the sampled drillcores. The clay content in sampled boreholes averaged 3.797 vol.% – a value that is similar to the synthetic orebody composition and comparable to our ML estimated clay content of 3.849 vol.% in this study. These minerals are variably distributed, and some are in contact with Cu-bearing minerals, which could further complicate mineral processing. Pre-

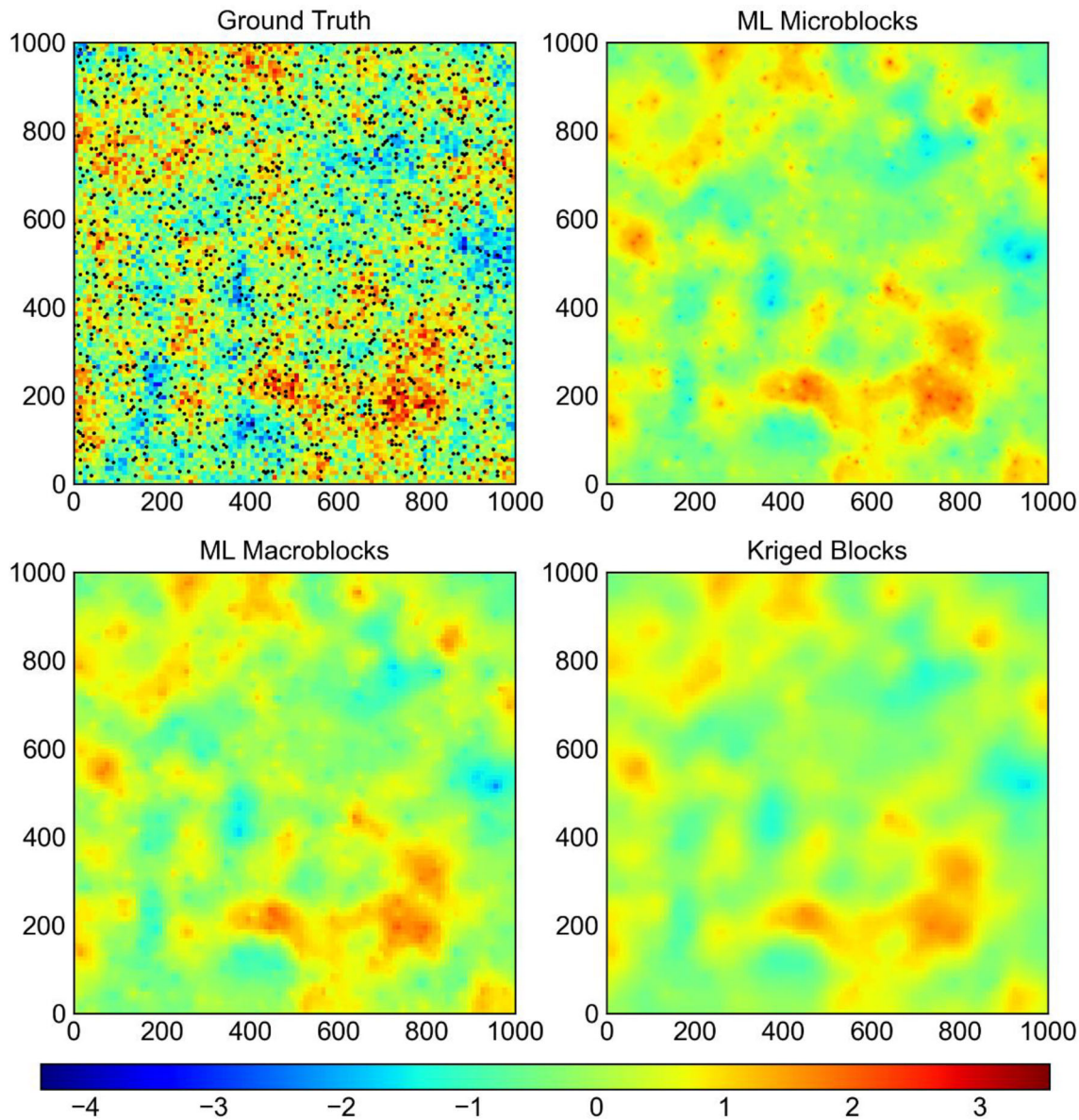

Figure 11. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under biased random sampling conditions. Here, the nugget corresponds to 0.65.

dictive geometallurgy of clay minerals can help understand their distribution and possible effects on mineral processing (Fig. 20). By incorporating mineralogical data into a block model, process engineers can identify areas with higher clay contents and modify their flotation reagent schemes accordingly. In practical geometallurgical studies, it is more important to model individual clay mineral contents (i.e., how much kaolinite, sericite, montmorillonite, etc.) as opposed to total clay content. This approach can also help the design of targeted testing programs to optimize flotation recovery in the presence of clay minerals. Mitigating clay mineral factors regarding flotation includes the use of dispersants and depressants to minimize clay interactions (i.e., this depends strongly on the clay mineralogy) with reagents and minerals, respectively. Also, modifications to the flotation circuit design, such as increasing the residence time or improving the froth stability, can help to mitigate the adverse effects of

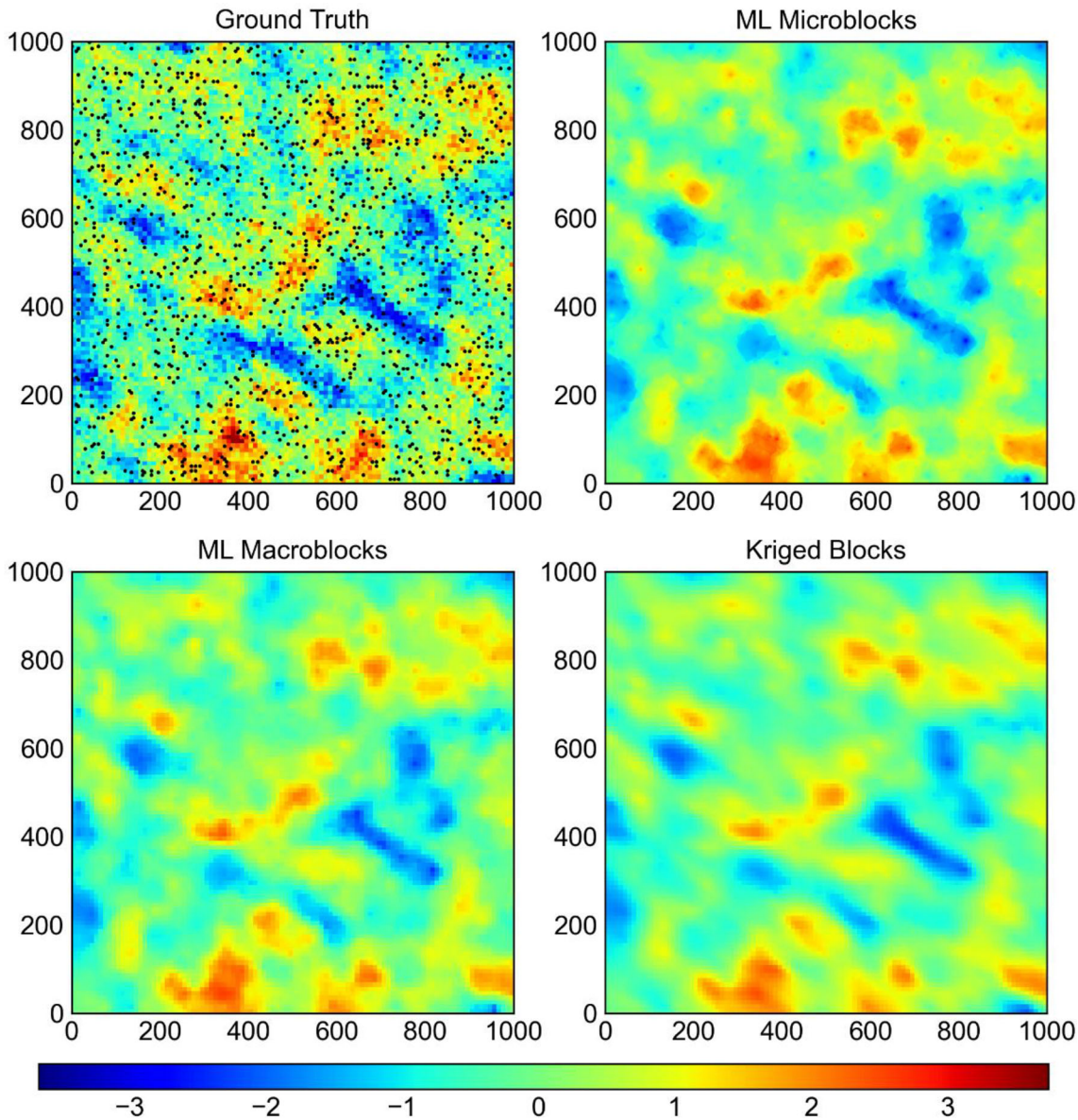

Figure 12. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under biased random sampling conditions. Here, the nugget corresponds to 0.1.

clay minerals on flotation recovery. The use of predictive geometallurgy to identify areas with high clay content can facilitate the design of targeted testing programs to optimize flotation recovery in the presence of these minerals (Fig. 20). For practical geometallurgical studies, a much higher resolution to capture the textural relationship between clays and sulfides will be required.

The BWI for the simulated porphyry Cu deposit ranged from 5 to 25 kWh/t_c. The average BWI of the sampled boreholes was 13.264 kWh/t_c, whereas the ML modeled value was 13.432 kWh/t_c and for the synthetic deposit, 13.210 kWh/t_c (Table 2). The BWI is affected by the mineralogical composition of the ore, as different minerals have different strengths, and therefore, responses to grinding (Fig. 21). For example, the presence of hard minerals such as quartz or pyrite can increase the BWI. In contrast, the presence of softer minerals such as clay minerals can decrease it. The BWI has important implications

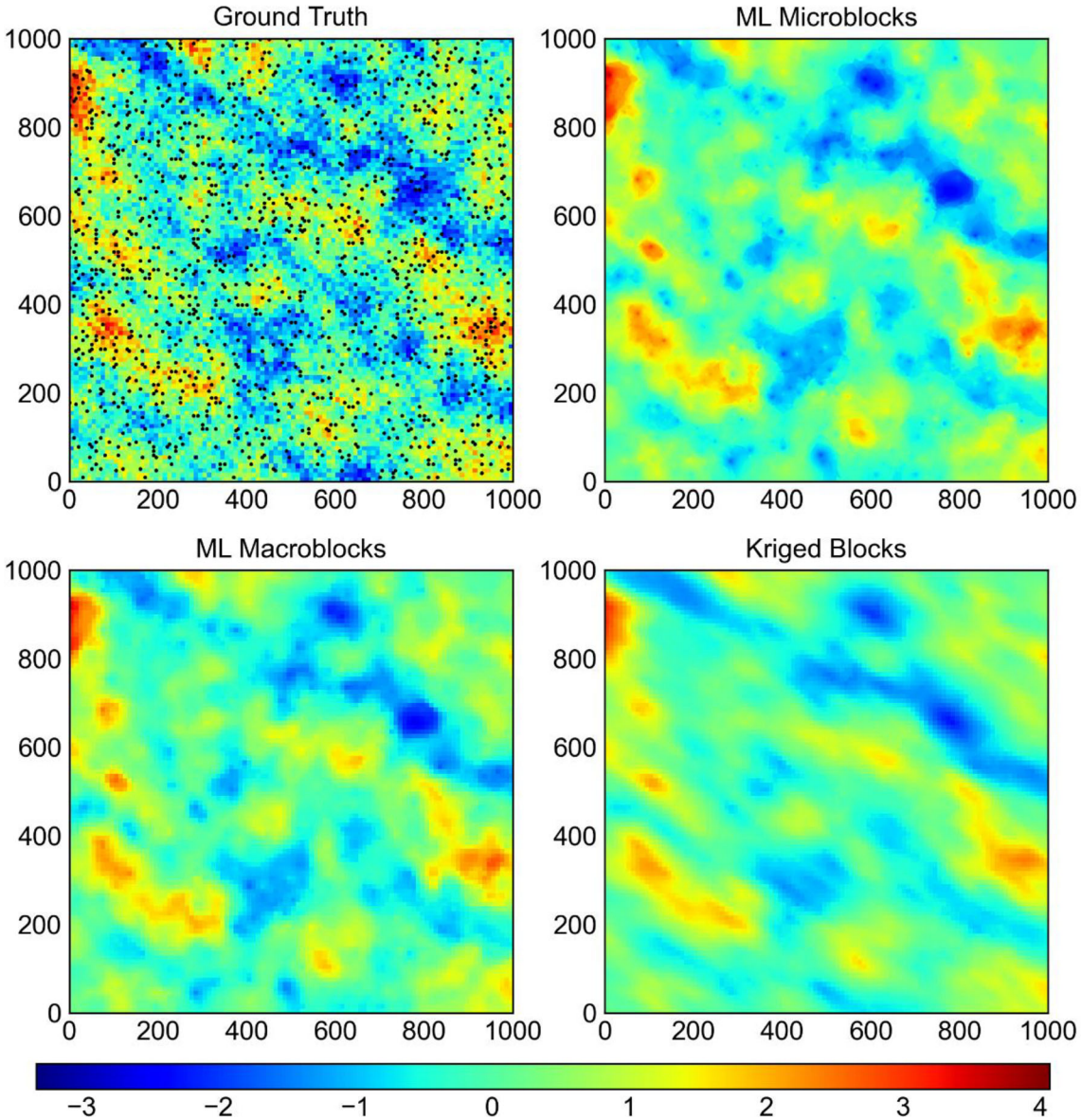

Figure 13. Ground truth (synthetic 2D block model) and results of the kNN (ML microblocks and macroblocks) and kriged modeling under biased random sampling conditions. Anisotropy corresponds to 0.1.

for the ore processing, as it affects the amount of energy required for grinding, which is a major contributor to the cost of ore processing. It also affects the durability of grinding equipment, as higher BWI values can result in greater wear on the grinding media and equipment. Predictive geometallurgy of the BWI is an important tool for understanding the properties of the ore and predicting expected energy costs. Through incorporating the BWI into a geometallurgical block model, it is possible to optimize the grinding circuit and reduce energy costs. Additionally, the BWI can be used to identify areas of the deposit with higher or lower energy requirements, allowing for more efficient mine planning and improved processing.

The rougher recovery of a porphyry Cu deposit is a critical metallurgical response variable that can significantly impact the economics of mining and

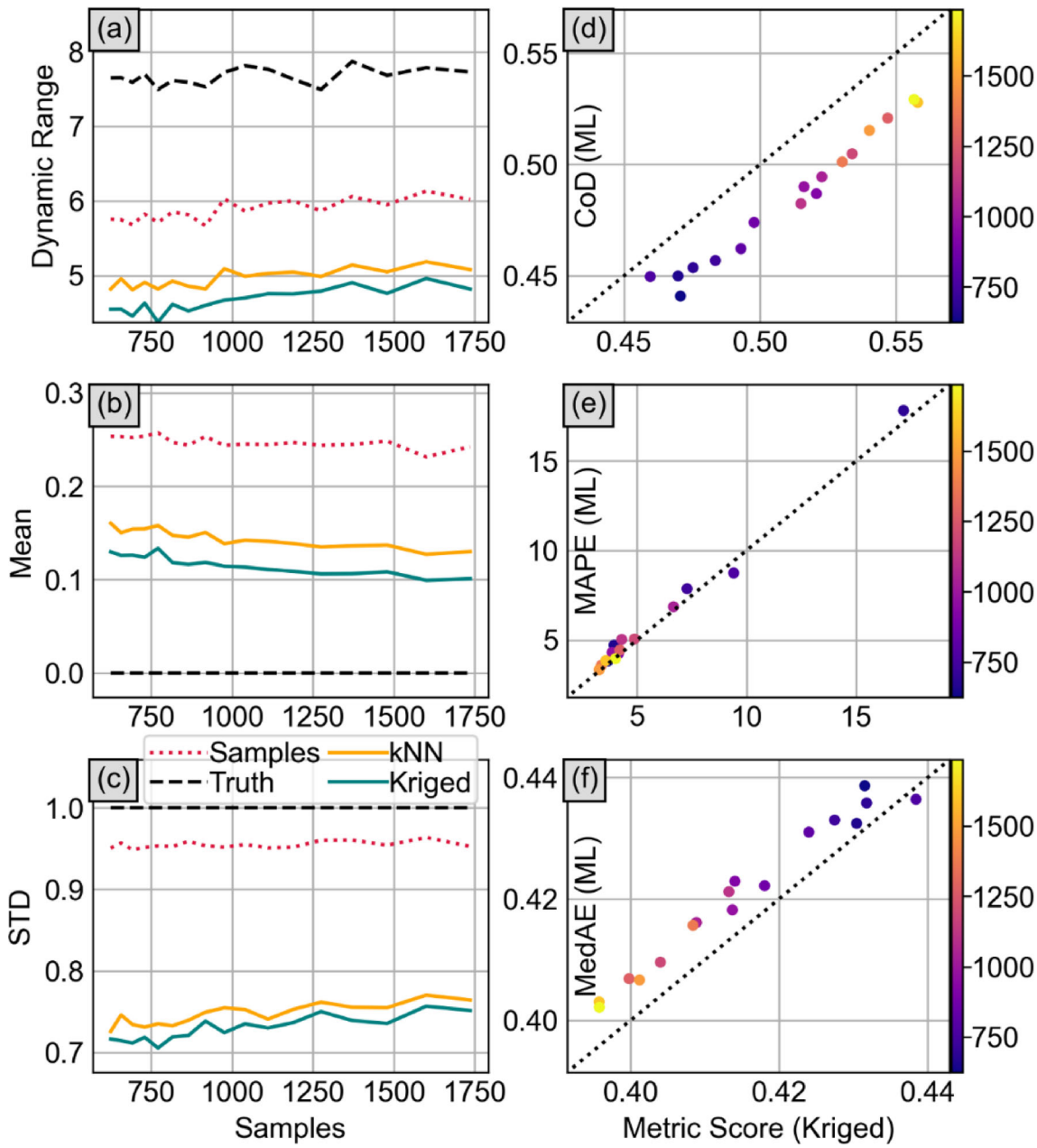

Figure 14. Parameter sweep of number of samples of biased sampling and impacts on (**a**) dynamic range, (**b**) CoD, (**c**) mean, (**d**), MAPE, (**e**) standard deviation, and (**f**) MedAE.

processing operations. In the present study, the rougher recovery results of the simulated porphyry Cu deposit indicated that the ore was amenable to metallurgical treatment with rougher recovery ranging from 75 to 95%. The range of recovery values was considered to be quite good for this type of deposit, which is primarily due to the favorable mineralogy of the deposit. Average rougher recovery of sampled boreholes was 83.384% while the actual recovery was 83.361%. These values are closely approximated in the estimate and average at 84.538% (Table 2). However, some areas of the deposit have lower recovery rates, i.e., less than 70%, which is attributed to the complex mineralogy

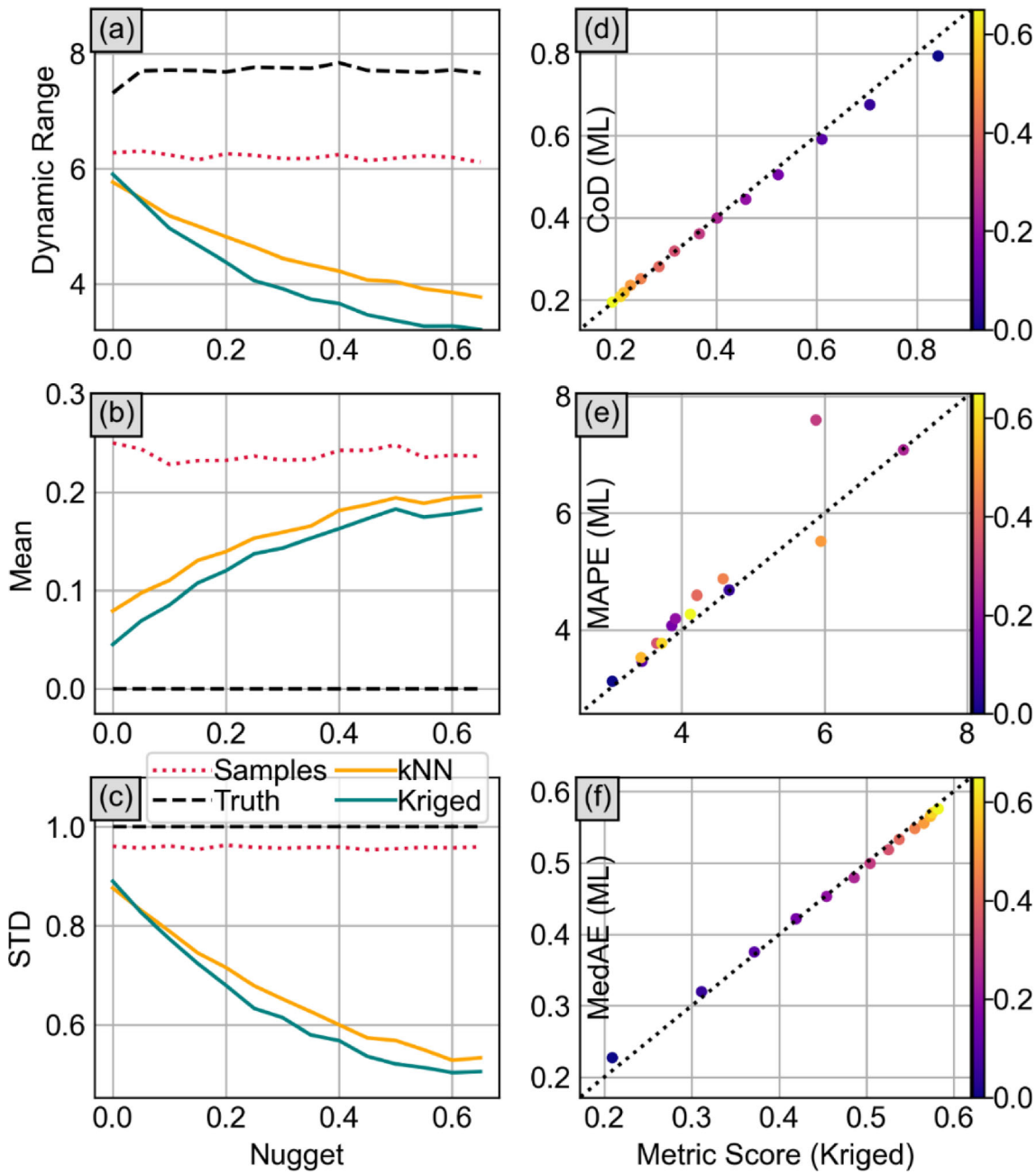

Figure 15. Parameter sweep of nugget effect of samples of biased sampling and impacts on (**a**) dynamic range, (**b**) CoD, (**c**) mean, (**d**), MAPE, (**e**) standard deviation, and (**f**) MedAE.

and alteration of the host rocks by clay minerals (Fig. 22). This indicates that the presence of clay minerals can negatively impact the recovery of valuable minerals during metallurgical processing. This underscores the importance of predictive geometallurgy in understanding the mineralogical composition and distribution of clay minerals in the ore, which can help in designing appropriate processing strategies to mitigate the adverse effects of such minerals. In addition, the rougher recovery response variable can also be used to predict ore properties and expected recoveries, which can, in

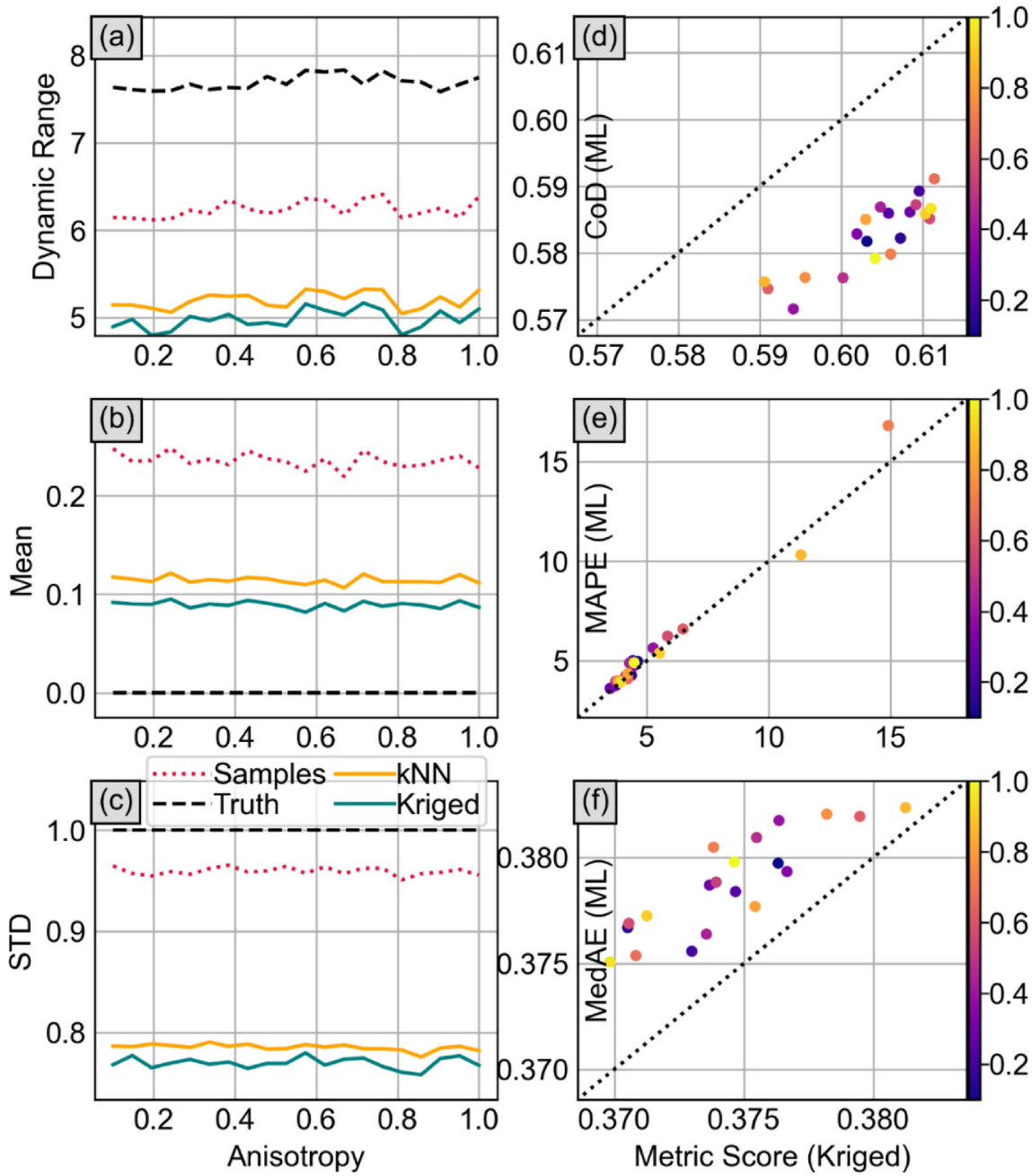

Figure 16. Parameter sweep of anisotropy of biased sampling and impacts on (**a**) dynamic range, (**b**) CoD, (**c**) mean, (**d**), MAPE, (**e**) standard deviation, and (**f**) MedAE.

turn, impact the selection of processing equipment, as well as the energy and cost requirements of processing. Therefore, incorporating a rougher recovery response variable into the block model can help optimize the entire mining value chain from exploration to processing.

DISCUSSION

Performance Summary and Implications

Block modeling is a common and important task in a variety of geoscientific, mining and related

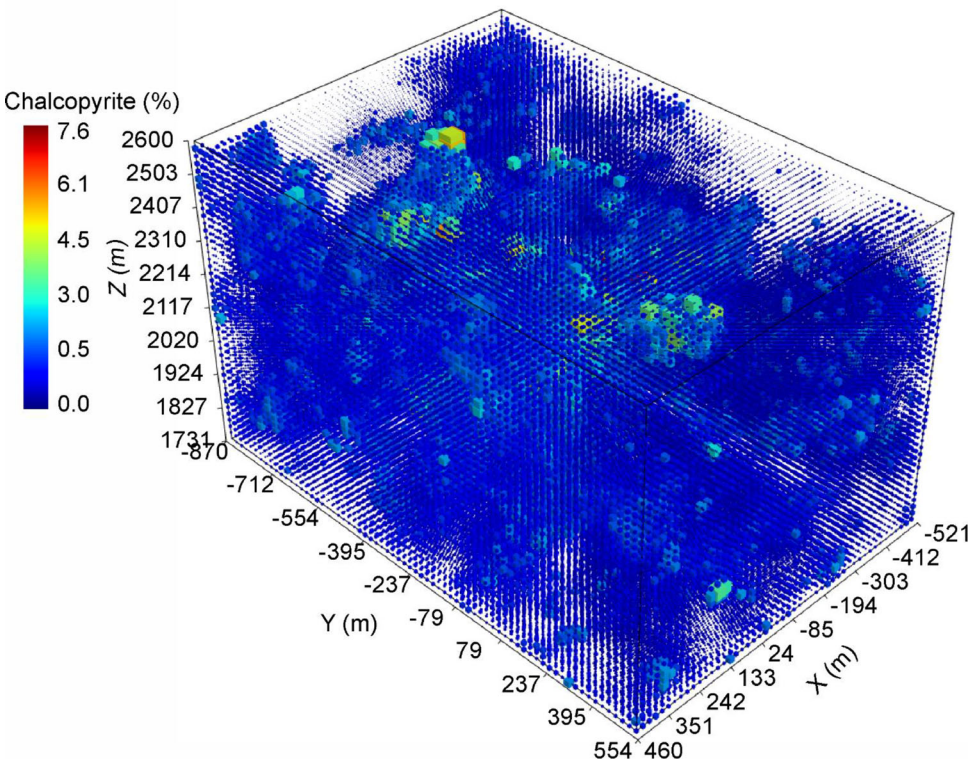

Figure 17. Predicted 3D block model showing the distribution of chalcopyrite (vol.%).

Table 2. Statistical quantities of the samples, predicted and actual 3D block models. The units for all minerals are in vol.% and for Cu, in wt.%

Materials	Mean	STD	Min	25%	50%	75%	Max
Chalcopyrite (samples)	0.736	0.541	0	0.411	0.648	0.917	7.430
Chalcopyrite (actual block model)	0.728	0.528	0	0.414	0.645	0.906	7.583
Chalcopyrite (predicted block model)	0.739	0.461	0	0.478	0.674	0.896	6.016
Bornite (samples)	0.124	0.151	0	0.023	0.084	0.166	2.292
Bornite (actual block model)	0.122	0.149	0	0.023	0.083	0.163	3.445
Bornite (predicted block model)	0.124	0.111	0	0.053	0.100	0.160	1.548
Cu (samples)	0.369	0.240	0	0.225	0.341	0.470	3.698
Cu (Actual block model)	0.364	0.236	0	0.225	0.338	0.462	4.060
Cu (predicted block model)	0.371	0.198	0	0.262	0.357	0.457	2.777
Clay minerals (samples)	3.797	2.897	0	1.827	3.087	5.078	40.670
Clay minerals (actual block model)	3.797	2.888	0	1.836	3.092	5.072	40.670
Clay minerals (predicted block model)	3.849	2.155	0	2.371	3.480	4.947	30.232
BWI (samples)	13.264	3.734	0	12.600	12.921	13.335	28.524
BWI (actual block model)	13.210	3.669	0	12.602	12.920	13.323	28.567
BWI (predicted block model)	13.432	2.888	0	12.650	12.914	13.324	27.896
Rougher recovery (samples)	83.384	16.947	0	84.197	87.358	89.399	94.368
Rougher recovery (actual block model)	83.361	17.008	0	84.178	87.343	89.383	94.857
Rougher recovery (predicted block model)	84.538	11.139	0	84.735	86.915	88.524	93.356

engineering activities. Block kriging is the dominant solution of the change-of-support problem by popularity in mapping and resource estimation. It uses a parametric model of the spatial variability structure to produce a weighted average. In comparison, current varieties of ML algorithms and, in particular,

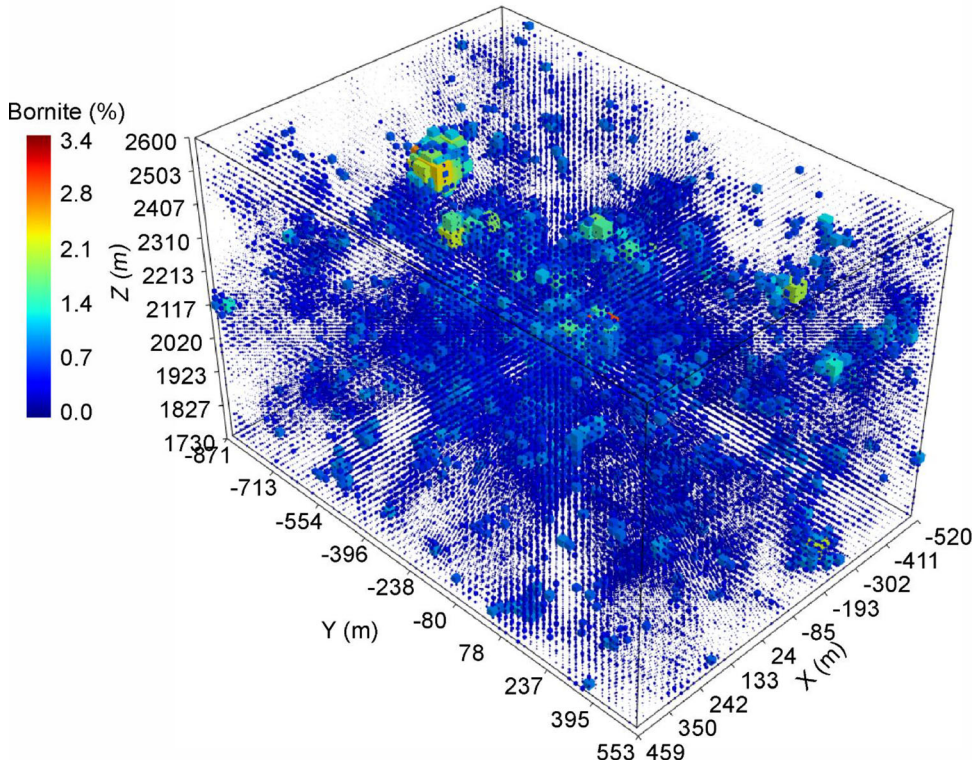

Figure 18. Predicted 3D block model showing the distribution of bornite (vol.%).

the kNN algorithm do not explicitly model spatial variability. Empirically, it is clear that under the conditions that were used for synthetic data generation and sampling, the performance of the kNN algorithm in our microblock-to-macroblock method varies depending on the specifics of the sampling practice and the metrics considered. Under regular sampling conditions, the kNN algorithm was more accurate at the mean to median level, as evidenced by the generally higher CoD, lower MAPE and MedAE scores (Figs. 6, 7, 8). However, the level of smoothing was generally lower for block kriging in comparison with the kNN algorithm (Figs. 6, 7, 8). This trend is reversed for results under biased sampling conditions (Figs. 14, 15, 16). This implies that under realistic conditions that are somewhere between completely regular and completely random sampling, there is no strong rationale to pick one method over the other for the purpose of block modeling. Within the explored conditions in this study, the performance differences as measured by the CoD, MAPE and MedAE metrics are generally small (typically on the order of 10^{-2} or a few percent absolute difference). To further understand the performance differences and implications, it is necessary to venture beyond empirical results and examine the algorithm architecture. As the present applicative overlap between ML and geostatistics is still small, we provide a discussion from two perspectives. The first perspective examines the algorithms in context of geostatistics, while the second perspective examines the algorithms in the context of ML.

The architecture of the kNN algorithm can be effectively examined through the perspective of geostatistics to understand the observed performance differences. The kNN algorithm uses the average of multiple neighbors, which in the case of prediction of each block within the microblock grid, means that the closest k-members of each microblock is used to derive an average. The idea of a neighborhood in ML is a generic one and only under the special circumstance that spatial coordinates are the sole features, is the concept of neighborhood identical to that in geostatistics. This implies that the support for each microblock is a fixed number of neighbors, which in the case of regular sampling, is a spatial neighborhood of fixed radius. In the case of

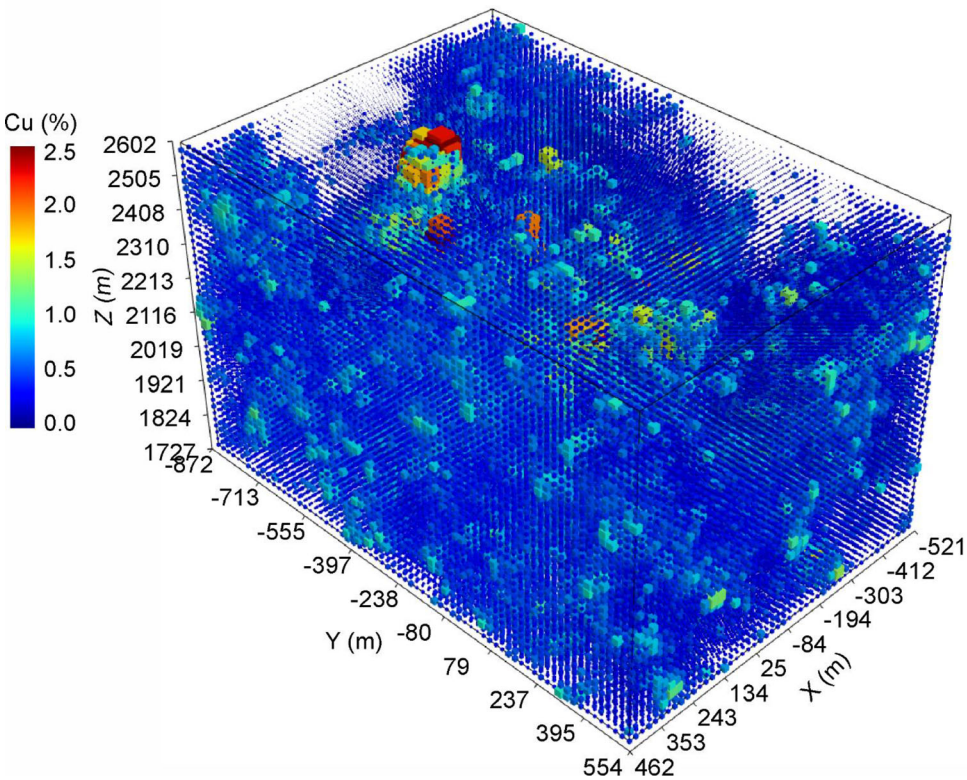

Figure 19. Predicted 3D block model showing the distribution of Cu (wt.%).

biased random sampling, this changes completely because the nearest neighbors do not occur at fixed distance intervals, and, therefore, the support is not a spatially consistent concept, and no fixed spatial neighborhood is applicable. In our experience, the biased random sampling is more representative of actual data, and hence, the concept of a fixed ''search'' radius in the case of block kriging does not generally have an equivalent in most ML algorithms. The way in which averaging of neighboring points is employed in kNN can technically be either unweighted or weighted. In this study, we used distance-weighting, which implies that contributions of points are inversely related to their distance. This is mechanistically similar to inverse distance weighting, which is a type of spatial modeling algorithm that is simpler than kriging. However, a key difference between kNN and inverse distance weighting is that kNN uses explicit control on the notion of the size of neighborhood through a single hyperparameter – the number of neighbors. This means that the delineation of the boundary of neighborhood is sharp in the case of kNN, which his generally not the case in inverse distance weighting. In other words, setting the hyperparameter ''k'' is equivalent to optimizing the number of neighbors to use in kriging, although in the latter, the model-tuning process is not necessarily cross-validation based. Fixing the number of neighbors through cross-validation implies that relative to inverse distance modeling, kNN is more resistant to over-smoothing (because of explicit controls on model bias and variance). In fact, the retrieval of fine-scale detail, particularly under more challenging conditions of biased random sampling, higher nugget effect and anisotropy (e.g., Figs. 11 and 13) is indicative of a reduction of reliance on distal supports.

The proposed microblocking approach using ML has demonstrated its effectiveness in reproducing ore distribution patterns and regularities in the simulated porphyry Cu deposit (Figs. 18 and 22). This approach was then used to create 3D predictive geometallurgical models that incorporate mineralogy, chemistry and metallurgical response variables, which can be used to improve mining, mine planning, grade control and mineral processing and

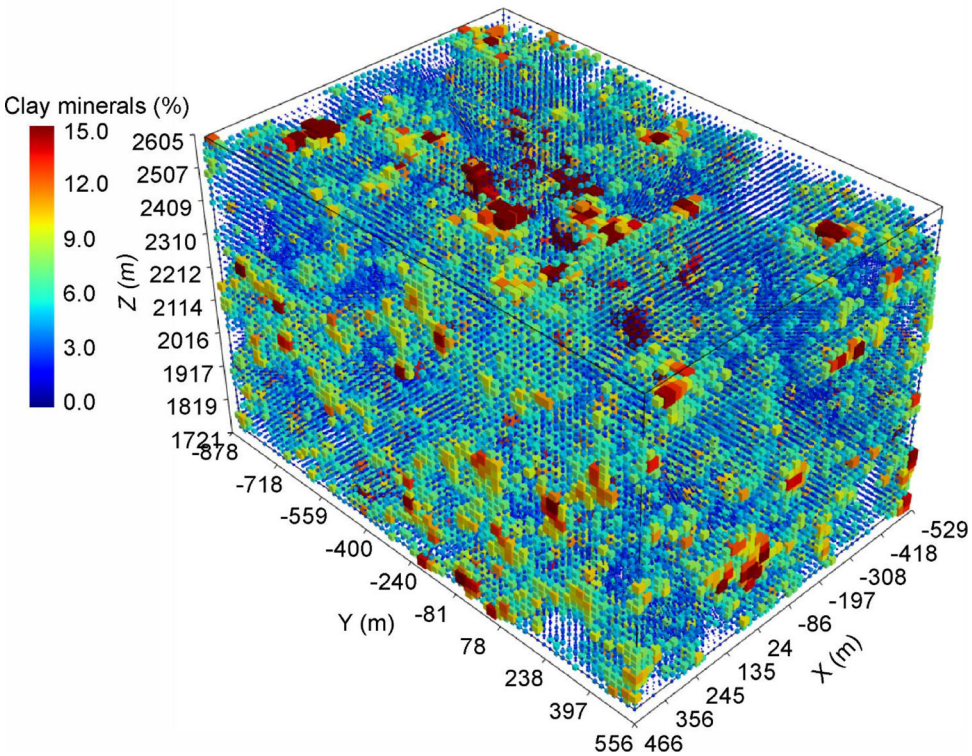

Figure 20. Predicted 3D block model showing the distribution of clay minerals (vol.%).

metallurgical extraction (e.g., Fig. 22). Using the microblocking approach, it was possible to accurately reproduce the concentration of the ore at sampled and block locations of the synthetic data during the deployment testing stages (Table 2). The predictive geometallurgical models generated using the microblocking approach provided valuable insights into the distribution of Cu, bornite, chalcopyrite and other minerals in the selected porphyry Cu deposit. These models also allowed for the identification of areas with complex mineralogy and alteration by clay minerals, which could result in lower recovery rates during metallurgical treatment. This information can be used to optimize mining, mine planning and grade control strategies to maximize the recovery of valuable minerals while minimizing costs. Moreover, the predictive geometallurgical models created using the microblocking approach can be used to optimize mineral processing and metallurgical extraction processes. By incorporating information about the mineralogy, chemistry and metallurgical response of the deposit, it is possible to design optimal processing circuits that maximize recovery while minimizing energy consumption and equipment wear.

From the ML perspective, the main task of a ML model is to capture as much information as possible that is contained within the given data. Block modeling is a highly specific application of ML in this sense, because in this case, the spatial information is captured by a ML model and replicated onto a fine microblock grid. This is similar to the classification application of domain boundary delineation using ML (Zhang et al., 2023). The performance of a ML model depends on its ability to model the relationships between the features and the data label contained within the data. In this sense, an appropriate choice of a ML algorithm should yield a minimum of bias and variance, such that the performance of the resulting models is maximized. For spatial learning tasks, the concept of local continuity of the spatial variability, or spatial correlation, is not explicitly addressed in known ML algorithms. In the case of the kNN algorithm, spatial learning is possible because the algorithm uses a notion of a local and weighted neighborhood combined with the use of distance metrics and spatial

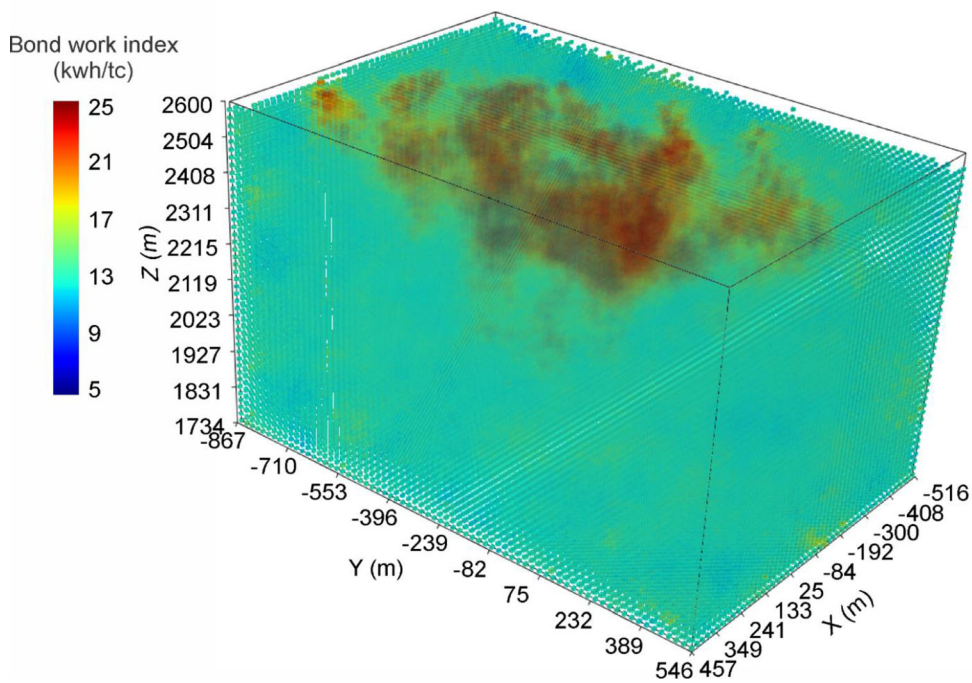

Figure 21. Predicted 3D block model of bond work index (BWI; kwh/t_c).

dimensions as features. Perfect replication of data is theoretically possible to within the constraints placed by sampling (e.g., resolution). As the data noise increases, for example, through increase of the nugget effect, the performance of models decreases (e.g., Figs. 7 and 15). This is best understood from the error decomposition into model bias, variance and irreducible error. Increasing the nugget effect increases irreducible error.

Despite the advantages of block kriging, including a very mature workflow and many application examples, there are some undesirable aspects of block kriging. One of the main challenges is its complicated and long sequence of procedures (Cressie, 1990). Furthermore, both ML-based and block kriging methods still require large amounts of data to produce accurate and reliable results, which can be a challenge in areas with sparse or limited data (Deutsch et al., 2014). Additionally, block kriging of various forms assumes some spatial statistical properties in the data (either global or at least satisfiable within geodomains, see Zhang et al., 2023), including that the geological variable being modeled is stationary. This assumption may not always hold true in practice, particularly in heterogeneous geological environments. Theoretically, this implies that methods that do not assume global data properties or require geodomaining could be generally more applicable, in that they require less considerations and treatise of the data prior to block modeling. Although this remains to be verified empirically. This is the case for the kNN algorithm at least in the realm of data science because we are unaware of any requirement to segment training data prior to the use of kNN-based classification or regression, although in specific cases (e.g., where there are highly heterogeneous clusters and prediction occurs within clusters), this is likely to be beneficial. Additionally, fitting of variograms and trimming of data tend to require expert judgment and substantial disciplinary experience. It is unclear what the net uncertainty across this sequence of procedures is because a systematic study has not been conducted. However, block models must always be used with grade control and metal reconciliation, which are feedback mechanisms to address model uncertainty (Chiles & Delfiner, 2012). The impacts of this type of uncertainty on mapping (e.g., for exploration, including prospectivity mapping) is wholly unknown, as reconciliation is not possible at this scale and for this type of activity. Consequently, the ability to minimize the length and complexity of

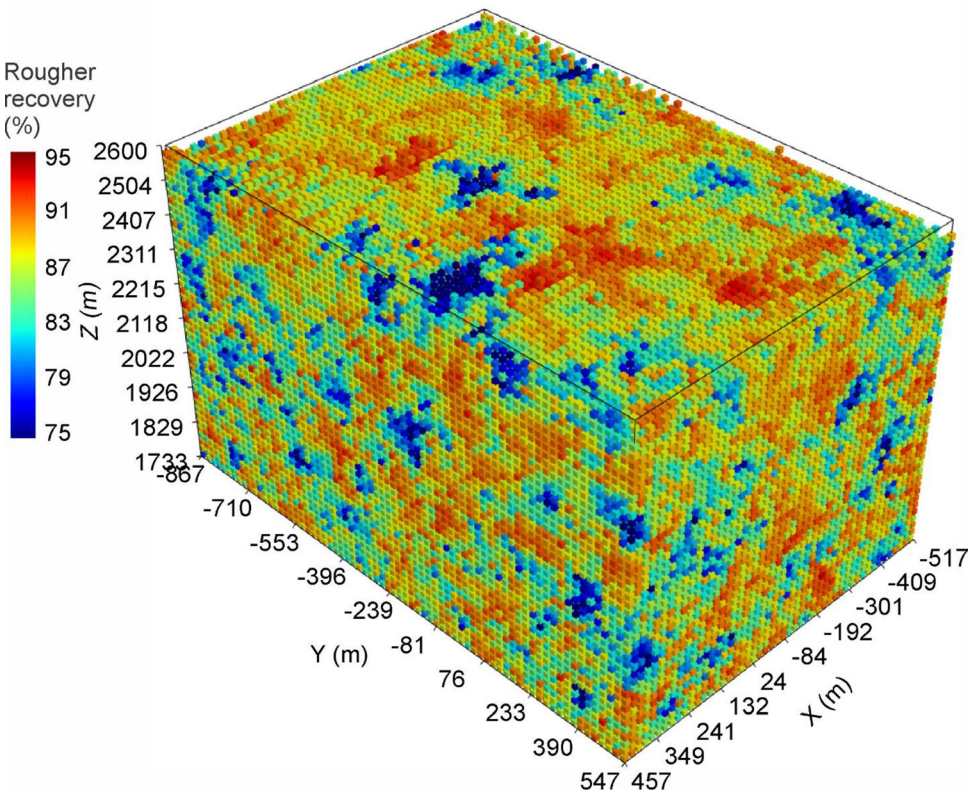

Figure 22. Predicted 3D block model of rougher recovery (%).

the sequence of procedures toward usable block models will likely increase model reproducibility and objectivity. In our case, the microblock-to-macroblock approach using the kNN algorithm makes use of a minimum of 2 key hyperparameters – the size of microblocks and the number of neighbors. The microblock size parameter is not as impactful on the outcome, as it can be set as small as computationally practical (e.g., as large grids consume more memory) and, in general, decreasing the parameter value results in increasingly diminishing returns. The tuning of the number of neighbors in kNN is straightforward and fully data-driven and automated using the ML workflow, which means no subjectivity is involved beyond data pre-processing (which was not employed in our study). This means that in comparison with block kriging, our method is essentially equally applicable to within the extent of data explored but is substantially simpler and essentially requires no manual contribution, aside from coding a workflow. This is an intended benefit of our approach, which enables data scientists to interpolate geospatial data using a machine learning-based workflow, which is consistent with their training. This benefit is likely to become more prominent with the rise of geodata science and higher-than-traditional velocity (ungridded) geospatial data.

CONCLUSIONS

Directly producing block models for a variety of purposes that include mapping and mineral resource estimation is challenging with ML algorithms because this task involves a change of support. ML algorithms implicitly assume that the prediction is made at the same size of support as the input and training data, and hence, unlike block kriging, changing support requires additional consideration. In this paper, we purposefully exploited the constant support property of ML algorithms and, in particular, explored the use of the simple kNN algorithm for the purpose of block modeling. We demonstrated that under the conditions of regular and biased random sampling, the kNN and microblock-

to-macroblock method yields results that are qualitatively and quantitatively similar to those of block kriging. In addition, detailed differences are nuanced in terms of qualitative and quantitative performance. We expect that our approach will be useful under a range of typical conditions, especially when highly reproducible and objective outcomes are expected, because our method requires tuning of a substantially reduced set of parameters as compared with the geostatistical approach. Our approach recognizes that the idea of punctual support is a mathematical approximation, which serves geostatistics well but is inconsequential to detrimental to ML in spatial tasks. Reconsidering this assumption and, therefore, undoing the punctual approximation enables the construction of a microblock grid, such that the individual predictions can be made at a scale that is much smaller than that of the final macroblock grid size and in a manner that is closer to the finite size of support as represented by the data. The two key recognitions are: (1) that the solution requires significant multidisciplinary knowledge at the geoscientific level (as in outside of the data science realm); and (2) modifying existing assumptions as necessary (notion of punctual support) to fit with ML algorithms. Formulating a fully automatable spatial interpolation method, with a workflow implemented in the ML framework would imply that our method should be highly accessible to data scientists, who may wish to venture into geosciences and perhaps, eventually become geodata scientists.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their comments which have greatly improved this manuscript. The authors also would like to thank E.J.M. Carranza for his generous time and editorial handling.

FUNDING

Open access funding provided by University of the Witwatersrand. Funding for this research through the Wits Mining Institute was provided by the DSI-National Research Foundation (NRF) Thuthuka Grant (Grant No: 121973) and DSI-NRF CIMERA. Additional funding for this research through Natural Resources Canada (Geological Survey of Canada) was provided by the Critical Minerals Geoscience Data (CMGD) program.

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

Abzalov, M. Z., & Humphreys, M. (2002). Resource estimation of structurally complex and discontinuous mineralization using non-linear geostatistics: case study of a mesothermal gold deposit in northern Canada. *Exploration and Mining Geology, 11*(1–4), 19–29.

Annels, A. E. (1991). *Mineral deposit evaluation, a practical approach*. Chapman and Hall.

Armstrong, M., & Champigny, N. (1989). A study on kriging small blocks. *CIM Bulletin, 82*, 128–133.

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

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

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). *API design for machine learning software: experiences from the scikit-learn project*. Retrieved May 13, 2023, from: http://arxiv.org/abs/1309.0238.

Carranza, E. J. M. (2009). Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and spatial association with geological features. *Ore Geology Reviews, 35*(3–4), 383–400.

Carvalho, D., & Deutsch, C. V. (2017). An overview of multiple indicator kriging. Geostatistics Lessons. Retrieved May 13, 2023, from: http://geostatisticslessons.com/lessons/mikoverview.

Cevik, I. S., Leuangthong, O., Cate, A., & Ortiz, J. M. (2021). On the use of machine learning for mineral resource classification. *Mining, Metallurgy & Exploration, 38*, 2055–2073.

Chiles, J., & Delfiner, P. (2012). *Geostatistics: Modelling spatial uncertainty* (2nd ed.). Wiley.

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

Cressie, N. (1990). The origins of kriging. *Mathematical Geology, 22*, 239–252.

Crochiere, R. E., & Rabiner, L. R. (1983). *Multirate digital signal processing* (vol. 18). Prentice-Hall.

David, M., (1976). The practice of kriging. In *Advanced Geostatistics in the Mining Industry: Proceedings of the NATO Advanced Study Institute held at the Istituto di Geologia Applicata of the University of Rome, Italy, 13–25 October 1975* (pp. 31-48). Springer Netherlands.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27–51.

Deutsch, J. L., & Deutsch, C. V. (2012). Kriging, stationary and optimal estimation: measures and suggestions. *CCG Annual Report 14*, Paper 306.

Deutsch, C. V., & Journel, A. G. (1992). *GSLIB: Geostatistical software library and user's guide*. Oxford University Press.

Deutsch, J., Palmer, K., Deutsch, C., Szymanski, J., & Etsell, T. (2016). Spatial modelling of geometallurgical properties: Techniques and a case study. *Natural Resources Research, 25*, 161–181.

Deutsch, J. L., Szymanski, J., & Deutsch, C. V. (2014). Checks and measures of performance for kriging estimates. *The Journal of the Southern African Institute of Mining and Metallurgy, 114*, 223–230.

Diaz-Gonzalez, F. A., Vuelvas, J., Correa, C. A., Vallejo, V. E., & Patino, D. (2022). Machine learning and remote sensing techniques applied to estimate soil indicators–review. *Ecological Indicators, 135*, 108517.

Dominy, S., O'Connor, L., Parbhakar-Fox, A., Glass, H., & Purevgerel, S. (2018). Geometallurgy: A route to more resilient mine operations. *Minerals, 8*(12), 560.

Dumakor-Dupey, N. K., & Arya, S. (2021). Machine learning – a review of applications in mineral resource estimation. *Energies, 14*(4), 4079.

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

Galetakis, M., Vasileiou, A., Rogdaki, A., Deligiorgis, V., & Raka, S. (2022). Estimation of mineral resources with machine learning techniques. *Materials Proceedings, 5*(1), 122.

Garrido, M., Sepúlveda, E., & Navarro, F. (2017). Optimisation of planning and scheduling of ore body with open pit extraction considering homogeneity in clays as geometallurgical variables. In *Geomin Mine planning, 5th International Seminar on Geology for the Mining Industry, 5th International Seminar on Mine Planning*, Santiago, Chile.

Garrido, M., Sepúlveda, E., Ortiz, J., Navarro, F., & Townley B. (2018). A methodology for the simulation of synthetic geometallurgical block models of porphyry ore bodies. In *Procemin GEOMET 2018, 14th International Mineral Processing Conference, 5th International Seminar on Geometallurgy*, Santiago, Chile.

Garrido, M., Ortiz J., Sepúlveda, E., Farfan, L., & Townley, B. (2019). An overview of good practices in the use of geometallurgy to support mining reserves in copper sulfides deposits. In *Procemin GEOMET 2019, 15th International Mineral Processing Conference, 6th International Seminar on Geometallurgy*, Santiago, Chile.

Garrido, M., Sepúlveda, E., Ortiz, J., & Townley, B. (2020). Simulation of synthetic exploration and geometallurgical database of porphyry copper deposits for educational purposes. *Natural Resources Research, 29*, 3527–3545.

Gelfand, A. E., Zhu, L., & Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics, 2*(1), 31–45.

Goertzel, B. (2006). *The hidden pattern: A patternist philosophy of mind*. BrownWalker Press.

Good, I. J. (1983). The philosophy of exploratory data analysis. *Philosophy of Science, 50*(2), 283–295.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press.

Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association, 97*(458), 632–648.

Groves, D. I., Vielreicher, R. M., Goldfarb, R. J., & Condie, K. C. (2005). Controls on the heterogeneous distribution of mineral deposits through time. *Geological Society, London, Special Publications, 248*, 71–101.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE, Montréal, Canada. https://doi.org/10.1109/ICDAR.1995.598994.

Hsieh, W. W. (2002). The impact of time-averaging on the detectability of nonlinear empirical relations. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 128*(583), 1609–1622.

Isaaks, E. (2005). The kriging oxymoron: a conditionally unbiased and accurate predictor (2nd Edition). In Leuangthong, O., Deutsch, C.V. (Eds.), Geostatistics Banff 2004. Quantitative Geology and Geostatistics, vol 14. Springer. https://doi.org/10.1007/978-1-4020-3610-1_37.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics*. Oxford University Press.

Jackson, J., McFarlane, A., & Olson, K. (2011) Geometallurgy – back to the future: scoping and communicating geomet programs. In *GeoMet 2011 – 1st AusIMM International Geometallurgy Conference 2011* (pp. 115–123). Australasian Institute of Mining and Metallurgy.

Kotsiantis, S. B. (2014). Integrating global and local application of naive bayes classifier. *International Arab Journal of Information Technology, 11*(3), 300–307.

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.

Krige, D. G. (1951). A Statistical approach to some mine valuations and allied problems at the Witwatersrand. M.Sc. thesis, University of Witwatersrand, Johannesburg, South Africa.

Krige, D. G. (1997). A practical analysis of the effects of spatial structure and of data available and accessed, on conditional biases in Ordinary Kriging. In E. Y. Baafi & N. A. Schofield (Eds.), *Geostatistics Wollongong 96. Quantitative Geology and Geostatistics* (Vol. 2, pp. 799–810). Kluwer Academic Press.

Krige, D. G., & Magri, E. J. (1982). Geostatistical case studies of the advantages of lognormal-de Wijsian kriging with mean for a base metal mine and a gold mine. *Journal of the International Association for Mathematical Geology, 14*, 547–555.

Kuipers, T. A. (2001). Structures in science: heuristic patterns based on cognitive structures. In *An Advanced Textbook in Neo-Classical Philosophy of Science* (vol. 301). Springer.

Lawley, C. J., McCafferty, A. E., Graham, G. E., Huston, D. L., Kelley, K. D., Czarnota, K., Paradis, S., Peter, J. M., Hay-

ward, N., Barlow, M., & Emsbo, P. (2022). Data–driven prospectivity modelling of sediment–hosted Zn–Pb mineral systems and their critical raw materials. *Ore Geology Reviews, 141*, 104635.

Lawley, C. J., Tschirhart, V., Smith, J. W., Pehrsson, S. J., Schetselaar, E. M., Schaeffer, A. J., Houlé, M. G., & Eglington, B. M. (2021). Prospectivity modelling of Canadian magmatic Ni (±Cu±Co±PGE) sulphide mineral systems. *Ore Geology Reviews, 132*, 103985.

Liu, Y., Carranza, E. J. M., & Xia, Q. (2022). Developments in quantitative assessment and modeling of mineral resource potential: an overview. *Natural Resources Research, 31*, 1825–1840.

Matheron, G. (1967). Kriging or polynomial interpolation procedures. *CIMM Transactions, 70*(1), 240–244.

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing, 39*(9), 2784–2817.

Mery, N., & Marcotte, D. (2022). Quantifying mineral resources and their uncertainty using two existing machine learning methods. *Mathematical Geosciences, 54*, 363–387.

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.

Olea, R. A. (1999). *Geostatistics for engineers and earth scientists*. Berlin: Springer.

Ortiz, J., Kracht, W., Townley, B., Lois, P., Cárdenas, E., Miranda, R., & Alvarez, M. (2015) Workflows in geometallurgical prediction: challenges and outlook. In *Proceedings of the 17th Annual Conference of the International Association for Mathematical Geosciences IAMG 2015*.

Ortiz, J. M., & Emery, X. (2006). Geostatistical estimation of mineral resources with soft geological boundaries a comparative study. *Journal of the South African Institute of Mining and Metallurgy, 106*, 577–584.

Parsa, M., Lentz, D. R., & Walker, J. A. (2023). Predictive modeling of prospectivity for VHMS mineral deposits, northeastern Bathurst mining camp, NB, Canada, using an ensemble regularization technique. *Natural Resources Research, 32*, 19–36.

Pyrcz, M. J., Jo, H., Kupenko, A., Liu, W., Gigliotti, A. E., Salomaki, T., & Santos, J. (2021). *Gestates python package*. PyPI, Python Package Index. Retrieved May 13, 2023, from: https://pypi.org/project/geostatspy/.

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

Richards, J. P. (2003). Tectono-magmatic precursors for porphyry Cu-(Mo-Au) deposit formation. *Economic Geology, 98*(8), 1515–1533.

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

Samson, M. (2020). *Mineral resource estimates with machine learning and geostatistics*. M.Sc. thesis, University of Alberta, Canada. https://doi.org/10.7939/r3-xxxz-5z86.

Sarma, D. D. (2009). Kriging Variance and kriging procedure. In *Geostatistics with applications in Earth Sciences* (pp. 125-138). Springer. https://doi.org/10.1007/978-1-4020-9380-7_8.

Srinivasan, K., & Fisher, D. (1995). Machine learning approaches to estimating software development effort. *IEEE Transactions on Software Engineering, 21*(2), 126–137.

Steiner, M. (2009). Empirical regularities in Wittgenstein's philosophy of mathematics. *Philosophia Mathematica, 17*(1), 1–34.

Talebi, H., Mueller, U., Tolodana Delgado, R., & van den Boofaart, K. G. (2019). Geostatistical simulation of geochemical compositions in the presence of multiple geological units: Application to mineral resource evaluation. *Mathematical Geosciences, 51*, 129–153.

Tosdal, R. M., & Richards, J. P. (2001). Magmatic and structural controls on the development of porphyry Cu±Mo±Au deposits. In: JP Richards, RM Tosdal (Eds.), *Structural controls on ore genesis, Reviews in Economic Geology* (vol. 14). Society of Economic Geologists.

Uddin, M. N., & Hamiduzzaman, M. (2009). The philosophy of science in social research. *The Journal of International Social Research, 2*(6), 1–11.

Vann, J., Jackson, S., & Bertoli, O. (2003). Quantitative kriging neighborhood analysis for the mining geologist – a description of the method with worked case examples. In *5th International Mining Geology Conference* (vol. 8, pp. 215-223). Bendigo, Australia. Melbourne: Australian Institute of Mining & Metallurgy.

Veronesi, F., & Schillaci, C. (2019). Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. *Ecological Indicators, 101*, 1032–1044.

Washburn, D. K., & Crowe, D. W. (1988). *Symmetries of culture: Theory and practice of plane pattern analysis*. University of Washington Press.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2nd ed.). Morgan Kaufman.

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*, 879–900.