

https://dspace.org/

Denoising of Geochemical Data using Deep LearningImplications for Regional Surveys

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

https://hdl.handle.net/10539/40828
WOS

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

Natural Resources Research, Vol. 33, No. 2, April 2024 (© 2024)
https://doi.org/10.1007/s11053-024-10317-5

Original Paper

Denoising of Geochemical Data using Deep Learning–Implications for Regional Surveys

Steven E. Zhang,[1,4] Julie E. Bourdeau,[1] Glen T. Nwaila,[3] Mohammad Parsa,[1] and Yousef Ghorbani[2]

Received 2 October 2023; accepted 20 January 2024
Published online: 28 February 2024

Regional geochemical surveys generate large amounts of data that can be used for a number of purposes such as to guide mineral exploration. Modern surveys are typically designed to permit quantification of data uncertainty through data quality metrics by using quality assurance and quality control (QA/QC) methods. However, these metrics, such as data accuracy and precision, are obtained through the data generation phase. Consequently, it is unclear how residual uncertainty in geochemical data can be minimized (denoised). This is a limitation to propagating uncertainty through downstream activities, particularly through complex models, which can result from the usage of artificial intelligence-based methods. This study aims to develop a deep learning-based method to examine and quantify uncertainty contained in geochemical survey data. Specifically, we demonstrate that: (1) autoencoders can reduce or modulate geochemical data uncertainty; (2) a reduction in uncertainty is observable in the spatial domain as a decrease of the nugget; and (3) a clear data reconstruction regime of the autoencoder can be identified that is strongly associated with data denoising, as opposed to the removal of useful events in data, such as meaningful geochemical anomalies. Our method to post-hoc denoising of geochemical data using deep learning is simple, clear and consistent, with the amount of denoising guided by highly interpretable metrics and existing frameworks of scientific data quality. Consequently, variably denoised data, as well as the original data, could be fed into a single downstream workflow (e.g., mapping, general data analysis or mineral prospectivity mapping), and the differences in the outcome can be subsequently quantified to propagate data uncertainty.

KEY WORDS: Deep learning, Geochemical surveys, Autoencoder, Denoise, Geochemical data.

INTRODUCTION

Geochemical surveys generate a sizable volume of consistent and often machine-readable data, which are a staple input for many downstream purposes, such as mapping, resource modeling, environmental monitoring and mineral prospectivity mapping (Grunsky & de Caritat, 2017; Roonwal, 2018; Sharpe et al., 2019; He et al., 2022). Generation of geochemical survey data begins with survey design, the purpose of which is to bias data coverage towards promising areas (if any, but depends on the setting and scale of survey) and create a spatial distribution of locations to be sampled, such that it maximizes the probability of geoscientific information retrieval (Garrett, 1983). A variety of sample

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

1520-7439/24/0400-0495/0 © 2024 Crown

media could be gathered and processed according to one or more established methodologies (e.g., Friske & Hornbrook, 1991; Dyer & Hamilton, 2007; Balaram & Subramanyam, 2022) and analyzed using various analytical methods (e.g., X-ray fluorescence [XRF], inductively coupled plasma-mass spectrometry [ICP-MS]). The key reason for strict adherence to methodology during the generation process is to ensure that the resulting data are accurate and precise (Balaram & Subramanyam, 2022). Precision is often assessed through comparing analyses of field and laboratory duplicates, while accuracy is assessed through the use of analytical standards (Geboy & Engle, 2011; Piercey, 2014). Despite a high level of rigor in generation methodology, geochemical data are not without uncertainty. Known sources of uncertainty include spatial or lithological coverage and coverage rate, and their variability and metrological errors (e.g., Sadeghi et al., 2021). For all physical systems being studied, there are intrinsic and extrinsic mechanisms that introduce uncertainty in geochemical data. Intrinsic reasons include the scale-invariance of complex systems, which is particularly applicable to mineral deposits, faults and river networks. This results in a constant multi-scaled statistical dispersion of any sampled quantity of a truly complex system. Other mechanisms for uncertainty related to methodology include an inability to exactly replicate procedures from sampling, preparation to analytical processes, and the fact that there is always some variability that is irreducible.

There are various ways to determine the level of uncertainty in geochemical data, and the most common approaches are used as part of the quality assurance and quality control (QA/QC) process (e.g., Knott et al., 1993; Geboy & Engle, 2011; Piercey, 2014; McCurdy & Garrett, 2016). Metrics of contamination, accuracy, precision and spatial variability are the most common. However, these metrics are used as part of the QA/QC process during the generation phase of the data. Subsequently, they are often used to infer the quality of data that do not feature actual ground truth (e.g., samples analyzed between reference materials). This creates two significant challenges in the modern era: (1) it is not obvious how the uncertainty of geochemical data can be determined in a post-hoc sense and how it can be minimized in general (e.g., denoised); and (2) uncertainty in the data cannot be generally propagated through complex models (e.g., machine learning [ML] models) and therefore the impacts of uncertainty are not generally anticipatable. These challenges are demonstrated by a paucity of publications revealing the relationship between traditional notions of contamination, accuracy, precision and spatial variability, and derived products, especially where artificial intelligence (AI)-based data modeling is used (e.g., prospectivity maps).

The downstream implications of geochemical data uncertainty are numerous and depend on the purpose of the data. Specifically, this issue is explicitly considered in resource estimation, where a poor spatial variability is known to be indicative of poor modeling outcomes that range from noisy to overly smoothed models (Rossi & Deutsch, 2013). Consequently, this is also the domain where geochemical data are met with the most rigorous application of spatial modeling, because modeling errors result in direct financial impact on exploration or mining companies. Therefore, it is standardized in the domain of resource estimation to publish all model parameters as part of an assessment report (Rossi & Deutsch, 2013; AusIMM, 2014). In the case of a rigorous application of geostatistics, several tasks are performed that include geodomaining, domain delineation, kriging neighborhood analysis, parameter tuning, model selection and post-modeling performance assessment (Rossi & Deutsch, 2013). Because geostatistics is highly explainable and many common varieties are linear, effects such as high nugget are easily appreciated and even anticipated by experienced geostatisticians (Isaaks & Srivastava, 1989; Clark, 2010). In particular, the relationship between uncertainty in input data and the resulting model is essentially known or predictable. In other domains outside of resource estimation, such as prospectivity mapping or traditional geochemical exploration, geostatistics is still used, but with generally under-documented workflows (see Grunsky, 2010, Grunsky & de Caritat, 2020 for examples of well-documented workflows). This creates additional uncertainty associated with workflow design in data analysis (e.g., Sadeghi et al., 2021). Consequently, prospectivity maps are almost never published with a geostatistical documentation of the data layers that were used, much less their effects on the resulting maps. It is unsurprising that it is generally not anticipatable the effects such as high nugget due to data noise, among other factors, on the resulting products. This is further complicated by the use of complex algorithms to model data, such as in AI-based prospectivity modeling. The use of unexplainable, or 'blackbox,' algorithms means that

the uncertainty of data cannot be analytically propagated to the downstream product. This is a serious scientific problem, because the accuracy and reproducibility of the products of blackbox-type approaches cannot be understood or be related to the scientific merit of the input data. More importantly, where data users cannot measure the quality of model outcome in relation to data input, there cannot be feedback to data generation processes based on objective criteria to evolve data generation to better support modern data usage scenarios, such as mineral prospectivity mapping.

Research on uncertainty propagation in AI-based methods in geosciences is currently in its infancy, and there are theoretically many approaches possible. As such, this study aims to understand the potential amount of uncertainty that is contained in regional geochemical surveys and its effect on mapping. In particular, we demonstrate that autoencoders, a type of artificial neural network, could be used to denoise geochemical data, which leads to a general increase in the spatial variability of the data, as observed through a reduction in the variogram nugget. In addition, we assess the effect of increasing levels of denoising and its relationship with the nugget. Lastly, we examine the relationship between the quality of the geochemical data and the effects of the autoencoder to deduce relationships that may be generally useful.

BACKGROUND AND METHODOLOGY

Geochemical Data

The geochemical data used in this study were generated from national-scale (Canada), multi-year (2010-2022) lake sediment survey programs (Figs. 1 and 2; Table 1). Based on the data characteristics, there are roughly two types of data available, legacy and modern. The legacy data generally contain elements analyzed and were generated prior to the 2000s. Extra material was retained as a survey practice to periodically re-analyze samples to cyclically modernize data (McCurdy et al., 2014). Consequently, the modern and legacy data overlap in spatial coverage and elements analyzed. This study only made use of the modern data, which contain a variety of elements analyzed and were constructed from amalgamating individual datasets (Table 1). In total, there were five independent datasets that cover portions of Manitoba, Saskatchewan and Labrador (Fig. 1). The original sampling occurred from 1976 to 1991 and was conducted by the Geological Survey of Canada as part of the national uranium reconnaissance (1975–1976), Manitoba mineral development agreement (1984–1989), and the exploration science and technology initiative (1991) programs (McCurdy et al., 2014). For details on the datasets, see Friske and Hornbrook (1991), Friske (1991), McCurdy et al. (2014), McCurdy and Garrett (2016), and Bourdeau and Dyer (2023). After amalgamation, the total number of data points that were not entirely empty was 11,302. The total abundance of the data varies per element analyzed (65 elements in total) and, for some elements, there were substantial amounts of missing data, which were a combination structural zeros and left censoring. All elements that featured more than 30% of the data missing were removed from subsequent usage (Fig. 3). Elements that were discarded included B, Ge, Hf, In, Pd, Pt, Re, Ta and Te, leaving 56 elements. The geochemical data were subsequently imputed using a k-nearest neighbor imputer, and the imputed data were transformed using the centered log-ratio transformation (CLR; Aitchison, 1986; Carranza, 2011), which is a type of data transformation that removes some of the undesirable geometric behaviors of compositional data (e.g., constant-sum limitation or closure).

Uncertainty in Geochemical Data

It is possible to contextualize uncertainty in geochemical data (Sadeghi et al., 2021) as a type of encoded variability. Some portion of the overall data variability is useful information from an information theory perspective. For example, for geochemical data that are intended to be used to discover mineral deposits through anomaly analysis (e.g., Grunsky & de Caritat, 2020), only variability that is spatially sensible and above the regional geochemical background can be considered contextually useful information (Sadeghi et al., 2021). Consequently, for this task, all variability below the regional background is essentially noise. For the purpose of the following discussion, we adopted a scheme to classify variability in geochemical data into several levels. The classes were: (1) data representation (coding or signal) level; (2) data content level; and (3) information (semantic or event) level. Class (1) variability is usually more relevant for streaming and complex data, such as audio-visual data, because such data

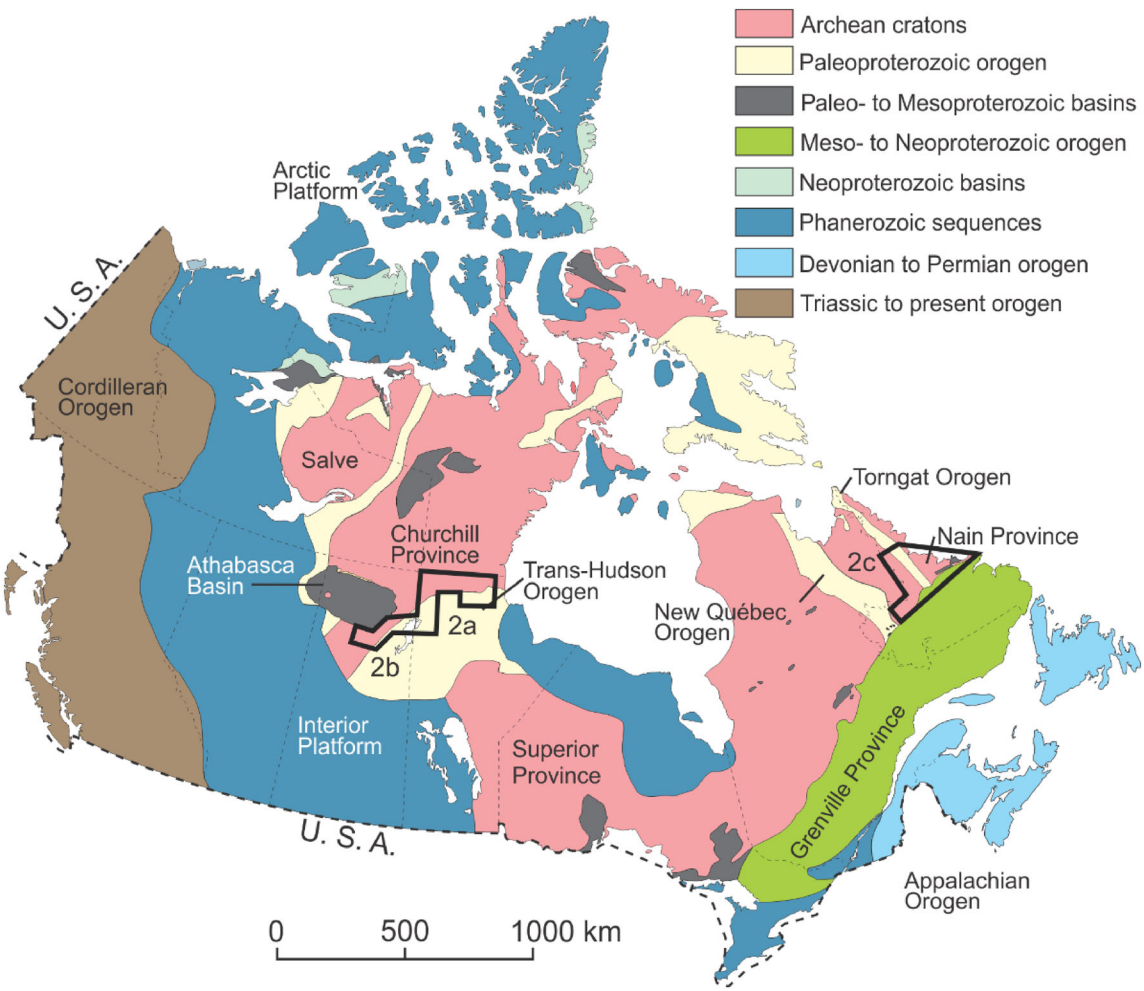


Figure 1. Simplified geological map of Canada with the location of the datasets in Manitoba, Labrador and Saskatchewan (black polygons). Detailed sample distributions, here labeled 2a, b and c, are presented in Figures 2a, b and c, respectively. Image taken and modified from Zhang et al. (2022).

are usually encoded and must be decoded to retrieve the original data. As such, the choice in coding (e.g., lossy encoding) and corruption in the signal (e.g., through data transfer) results in changes in the decoded data. This class of variability is rare in typical geochemical data after analytical instrumentation, because such data are not encoded beyond digital representation and is already reduced from time-varying signals through integration of spectra using either fluorescence (e.g., XRF) or mass spectrometry (e.g., ICP-MS) techniques. Class (2) variability is significant in geochemical data and assuming that database corruption is not a factor; then, the remainder is associated mainly with uncertainty of geochemical data. Uncertainty causes measured values to fluctuate upon successive re-measurements (with the physical system unaltered in between). In this view, inaccuracy and imprecision, for example, are both noise added to perfectly accurate and precise data. Class (3) variability is induced by physical (chemical) variability of the natural system that is captured by the data and, in cases of data intended for mapping, should be the main source of data variability. This study is concerned with the modulation of class (2) data variability.

There are two domains of uncertainty in geochemical data (e.g., Shanahan et al., 2013): (1) spatial; and (2) variable. Spatial uncertainty, which pertains to the measurement error associated with the coordinates and the punctuality approximation of samples, contains two major sources of error–uncertainty of the position measurement (e.g., the

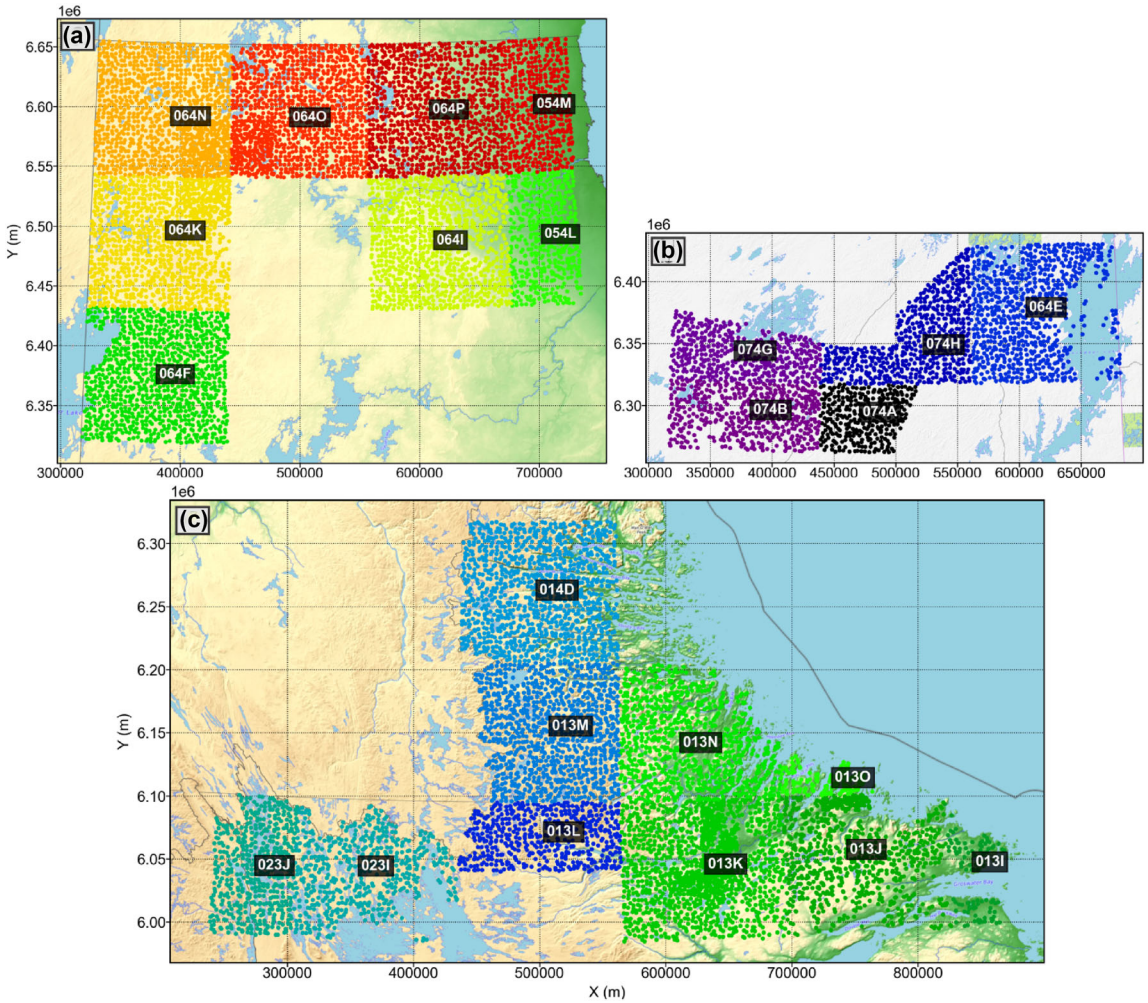


Figure 2. Detailed sample locations in (**a**) Manitoba, (**b**) Saskatchewan and (**c**) Labrador. The positions of the samples within Canada are presented in Figure 1.

uncertainty of a global positioning system or GPS) and uncertainty introduced by natural processes, such as spatially varying strengths of mixing and transport. The first type of measurement error was greater prior to the use of satellite-based positioning systems and would be the case for surficial geochemical samples that were gathered before the widespread use of GPS. This is the case for at least some samples in our dataset, because the re-analyses did not re-assess the spatial coordinates of the original samples (which would be impossible without re-sampling). Even for data that were gathered with the aid of satellite-based positioning systems, the spatial accuracy of such data was inconsistent because of changes in the capabilities of the positioning systems. Spatial uncertainty at the point-scale appears as noise added to the coordinates of the samples, whose bandwidth and magnitude depend on the uncertainty characteristics of the position measurements. At the regional scale, spatial uncertainty is insignificant in comparison with spacing between samples (roughly 1 sample per 6–13 km^2; Friske, 1991; Friske & Hornbrook, 1991). This is not true for resource-scale subsurface sampling, because it is not directly benefited by satellite-based positioning systems (Davis, 2010). In relative terms (e.g., using the relative error metric), spatial uncertainty is also smaller than aspatial uncertainty. This study focuses on variable domain (aspatial) uncertainty in geochemical data. This type of uncertainty is typi-

Table 1. Details of the geochemical datasets used in this study. Samples for modern datasets were digested using aqua regia followed by ICP-MS analysis.

Province	Zone	Year	Open File	Samples	Open File re-analysis	Elements
Manitoba	64N, O, K	2013	7309	1198	321, 322, 323	Ag, Al, As, Au, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Gd, Ge, Hf, Ho, In, K, La, Li, Lu, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Pr, Rb, Re, S, Sb, Sc, Se, Sm, Sn, Sr, Ta, Tb, Te, Th, Ti, Tl, Tm, U, V, W, Y, Yb, Zn, Zr
	64I, P, 54L, M	2010	6671	806	407, 408	Ag, Al, As, Au, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Fe, Ga, Ge, Hf, Hg, In, K, La, Li, Mg, Mn, Mo, Na, Nb, Ni, P, Pb, Pd, Pt, Rb, Re, S, Sb, Sc, Se, Sn, Sr, Ta, Te, Th, Ti, Tl, U, V, W, Y, Zn, Zr
	64F	2023	8948	966	1104	Ag, Al, As, Au, B, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cs, Cu, Dy, Er, Eu, Fe, Ga, Gd, Ge, Hf, Hg, Ho, In, K, La, Li, Lu, Mg, Mn, Mo, Na, Nb, Nd, Ni, P, Pb, Pd, Pr, Pt, Rb, Re, S, Sb, Sc, Se, Sm, Sn, Sr, Ta, Tb, Te, Th, Ti, Tl, Tm, U, V, W, Y, Yb, Zn, Zr
Saskatchewan	74A, B, G, H	2021	8837	1290	1359	
	64E, 74A, H	2022	8871	1179	1643	
Labrador	13L, M, 14D, 23I, J	2016	8026	3441	557, 559, 560, 904	
	13I, J, K, N, O	2023	8962	3163	513, 558, 997, 1636, 1637, 2645, 2646, 2648, 2650	

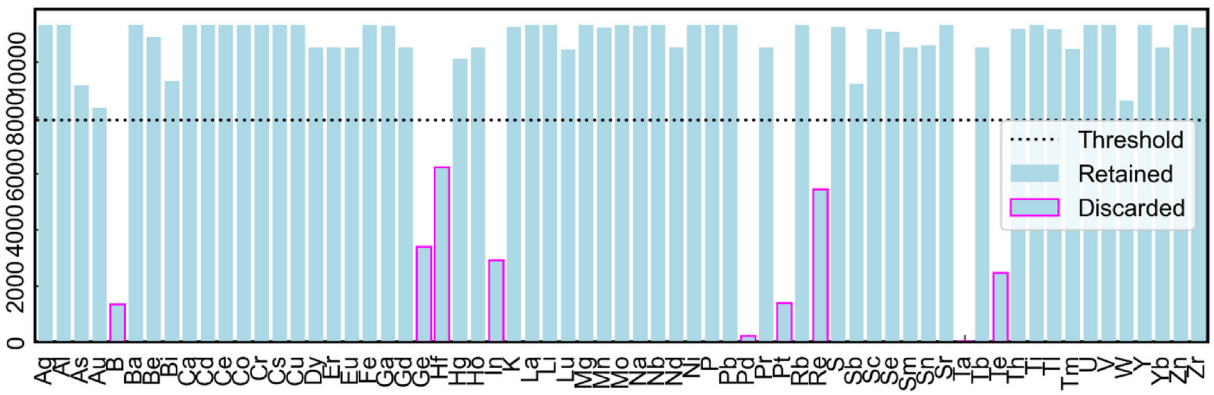


Figure 3. Abundance of non-empty records per element in the dataset (65 elements total). Data for elements that do not meet the threshold (30% missing values) were discarded.

cally decomposed into contamination, accuracy, precision and measures of spatial variability (e.g., Geboy & Engle, 2011; Piercey, 2014; McCurdy & Garrett, 2016). It appears as multivariate correlated noise added to perfectly accurate and precise data. Although it is not known whether the nugget model or the random noise model better characterizes the uncertainty in geochemical data in general, both are applicable and can result in similar outcomes. For sediment geochemical data, sediment mixing through transport and the fine particle sizes relative to the size of samples essentially create a smoothing effect to any underlying distribution function. Consequently, for finite (non-repetitive) measurements in an area, the geostatistical nugget captures effects induced by random noise. In the case of this study, the field duplicates generally demonstrate a low (inferred) impact of the nugget model, as revealed through ANOVA (e.g., McCurdy et al., 2013). Within the random noise model, perfectly certain data are perfectly correlated between samples taken with zero distance between them, which implies that the spatial variability is maximized, and the nugget effect is zero. In the case of the data used in this study, the QA/QC metadata for all datasets documented the usage of reference material standards to

assess accuracy, analytical duplicates to assess precision, and field duplicates to assess spatial variability through ANOVA at two scales (in-group and between-group scales) (e.g., McCurdy et al., 2013).

There are many known detailed mechanisms that could result in a measurable loss of accuracy, precision and spatial variability that roughly divides along the data generation pipeline. During sampling, variations in the sampling process and the representativity of the sampling volume can result in increases in data uncertainty (Govett, 1983). For these reasons, sampling processes are often standardized, and a protocol is used to guide activities (e.g., Knight et al., 2021; Bourdeau & Dyer, 2023). There are, however, unmeasurable, and essentially unknowable variabilities in the execution of any standardized operating procedure, and geochemical sampling is no exception. Similarly, in the laboratory, instrument drift, operator consistency, procedural rigor and other aspects impart additional uncertainty (e.g., Piercey, 2014). These types of variability are extrinsic to natural processes being studied and occur anywhere along the data generation process. The intrinsic sources of variability due to natural processes can include (Crock & Lamothe, 2011; Balaram & Subramanyam, 2022): the existence of refractory minerals, which especially impact incomplete digestion-based methods (e.g., aqua regia digestion), such as zircon, chromite, rutile and barite; spatial heterogeneity; and matrix effects. The effect of spatial heterogeneity at the scale of sampling, processing, and analysis introduces the nugget effect, which is observable as a degradation of spatial correlation at short ranges. This lowers measures of spatial variability. The strength of the nugget is geostatistically the most significant important spatial metric for assessing spatial variability (Isaaks & Srivastava, 1989; Rossi & Deutsch, 2013).

Review of Denoising of Geochemical Data

Methods to denoise geochemical data, which is equivalent to the selective removal of class (2) variability, are relatively uncommon and appears to be an emerging domain of research. It is important to recognize that this problem is an ill-posed inverse problem, which means that there are no unique solutions (Fan et al., 2019). The most well-developed sub-domain appears to be the denoising of gridded geochemical data, because this sub-domain benefits directly from extensive development of the broader domain of geostatistics and image denoising (Fan et al., 2019). An application-specific method based on geostatistics and spatial filtering methods to denoise geochemical data employed Gaussian transforms and simulations for the purpose of improving the performance of classification tasks downstream (Guartán & Emery, 2021). This method is application specific because its output is optimized for tree-based classification algorithms. Geostatistical simulations are pioneered on geospatial data, but spatial filtering is a classic image denoising method (Fan et al., 2019). It was recognized that denoising geochemical data improved short-range correlation, which improved downstream modeling outcomes (Guartán & Emery, 2021). In particular, they observed that the nugget effect was reduced, which they interpreted to be a reduction in unknown noise that is influenced by sampling density and grid design, as well as by the dynamical complexity of the environment, analytical errors or other site-specific variation (Carrasco, 2010; Hofmann et al., 2010). However, it is worthwhile to highlight that both nugget and random effect models contribute to the strength of the nugget effect. In particular, a physical process affected by pure nugget effect cannot yield consensus under repeated sampling that differs by space, as the underlying function is spatially discontinuous. Consequently, it is possible to decrease the nugget effect, but only up to the limit of the total suppression of random noise in spatial data. Geostatistical simulation-based methods suffer from a key unique issue, that back transformation is generally impossible using filtered simulations (Guartán & Emery, 2021), which prevents general uses of denoised data.

A different class of approaches adopts pure signal processing techniques (Fan et al., 2019). Insomuch as geochemical data can be suitably regarded as image data (in gridded form), all signal processing techniques are theoretically applicable, but application examples are rare. This class contains classical denoising methods, such as: spatial filers (e.g., moving average filters or the Wiener filter); variational methods; regularization-based methods; sparse representation-based methods; and low-rank minimization-based methods (Fan et al., 2019 and references therein). Although spatial filters are common in GIS applications, especially in imagery or remote sensing tasks, they are not well documented in the task of denoising of specifically geochemical data. This class also contains transform domain-based methods, such as: data adaptive (e.g.,

independent component analysis); non-data adaptive (spatial-frequency or wavelet domain-based); and collaborative filtering (Fan et al., 2019). With the exception of wavelet-based methods, we are unaware of any documented examples of the application of other transform domain-based methods to geochemical data. Geochemical maps can be transformed from the spatial into the wavelet domain, which decomposes the original signal into a scale-space representation. Similar to the frequency domain (via a Fourier transform), scales can be selected or attenuated explicitly, which enables the targeting of noisy signatures (e.g., short-range noise) (Liu et al., 2014). The cleaned data that are transformed back into the spatial domain can exhibit increased correlation at short ranges, if high-frequency noise was removed.

The most popular ML-based approaches to denoise gridded data rely on convolutional autoencoders (Fan et al., 2019; Ilesanmi & Ilesanmi, 2021). In image denoising tasks, convolutional autoencoders are considered the state-of-the-art (Fan et al., 2019). Indeed, treating noises with a significant spatial signature have a substantial history in image processing, such as the removal of image-compression artifacts (Teng et al., 2023). In the deep learning approach, embedding an autoencoder into a convolutional neural network creates a general-purpose anomaly detector or remover for spatially gridded data (Fan et al., 2019). Although anomalies can include data noise, the intention in geospatial anomaly detection is always to target for scientifically or economically relevant entities (e.g., Guan et al., 2021). This process relies on data reconstruction at the image level, which progressively removes class (3) variability until some stopping criterion has been satisfied. In doing so, class (2) variability is removed automatically, as they typically contain short-range variability that is smaller than class (3) variability in high-quality data (e.g., a deposit size variation in data versus pair-wise fluctuations in measurement). Consequently, anomaly detection of sufficiently large geochemical anomalies de facto involves denoising of geochemical data. However, a key difference is in the stopping criterion. To denoise data, it is important to preserve most if not all of class (3) variability, which implies that the extent of information loss encountered in anomaly detection centered around geoscientific anomalies is excessive for strictly the removal of class (2) variability. However, this is desirable in some uses of data, such as in prospectivity mapping, where modulation of class (3) variability is the AI task and noise is (generally) detrimental to the outcome anyway. Convolutional autoencoders and data reconstruction is the combination employed in a variety of geoscientific anomaly detection studies to modulate class (3) variability (e.g., Guan et al., 2021; Xiong & Zuo, 2021; Zhao et al., 2023). The modulation of class (3) variability has received more attention, probably because it constitutes data usage as compared with the latter, which can be part of data generation or processing but has direct impacts on usage. It is important to note that the algorithm choice and architecture are dependent on the task, and there is no standardized architectural design (Fan et al., 2019), and for the task of mineral prospectivity mapping using spatial anomaly detection, gridding is unavoidable. However, denoising functionality does not originate from gridding or the interface of deep learning architectures to gridded data through convolution, which is an image processing technique to extract spatial features, such as edges. Rather, denoising occurs from data reconstruction via the autoencoder that is embedded deeper in the architecture. The convolution stages extract spatial features, which are then compressed with information loss through the autoencoder. Deep learning-based approaches are theoretically as applicable (if not more so, because of the generality of neural nets) as both geostatistics and signal processing-based denoisers. However, the methods summarized herein suffer from a notable weakness, which is the requirement of the gridding of geochemical data.

In geochemical exploration, it is rare to encounter primary data that are natively gridded. Notable exceptions include secondary (inverted) geochemical data from remote sensing data, which produce a form of fully gridded data (Zhang et al., 2023a, b). It is also common to encounter aspatial uses of geochemical data, such as in: data analysis (e.g., Grunsky & de Caritat, 2020); theoretical mineralogy and petrology (e.g., Pearce, 1968); applicative mineralogy (e.g., Ulrich et al., 2021); and environmental assessments (e.g., Rodushkin et al., 2022). Additionally, geochemical data can either contain no spatial coordinates or are unsuitable for gridding; for example, geochemical data describing similar rock types without explicit spatial continuity (e.g., occurrence of igneous rocks of the Karoo Supergroup, see Zhang et al., 2021b). Spatial interpolation must be applied to ungridded data to create fully gridded data. Ironically, data noise degrades interpolation performance (e.g., increased

nugget effect decreases spatial correlatability). Consequently, denoising methods that require interpolate-then-denoise are not the most general purpose by design, because they require gridding (effectively the creation of maps), which is a data usage task that must occur on noisy data, implying that data must be used in its noisiest state prior to treatment. In addition, interpolating data first (gridding) introduces additional variability into data, by altering noise variances, such as through the volume-variance effect. For example, over-smoothing data during interpolation are equivalent to pre-application of spatial or frequency filters. This is compounded by the fact that interpolation standards exist universally only in resource assessment (OSC, 2011; SAMREC, 2016). Trained deep learning models are data-dependent and cannot be deployed on differently interpolated or compressed data without re-training (Fan et al., 2019). Using interpolated data would also create challenges with directly relating the denoised data to the original in terms of its quality (e.g., precision and accuracy) as assessed during data generation. This is because gridding introduces a change of support, which cannot be applied to quality metric results that were derived on a per-sample basis (Nwaila et al., 2023). Consequently, gridding of sparse data adds additional complexity to understanding the effect of data quality on downstream uses.

A lack of generality is not a problem for task-specific denoising of data, because unlike pure research, in applicative or engineering domains, the outcome is in the form of a specific and not general solution of the problem. However, this is an issue in method development, because a general method is more desirable, can serve as a functional baseline for future development, and is more applicable to a larger diversity of data and downstream tasks. It is clear that it would be preferable to denoise first, then interpolate, because it is the most general solution and even interpolation itself could benefit from the use of cleaner data. This is essential toward understanding how data uncertainty can propagate through any model, by isolating and quantifying the contribution of each type of uncertainty. A general approach to denoising geochemical data without using spatial coordinates is theoretically possible using deep learning algorithms. In addition, this type of approach would avoid the issues associated with back transformation, and changes to data including those incurred through the gridding process. A caveat with techniques that use ungridded data natively is that they treat data in the variable domain, which implies that only multivariate data could be denoised, whereas image processing techniques or geostatistical methods are usable on univariate data (e.g., gray scale imagery).

Autoencoders and Task Formulation

Classic autoencoders are a type of unsupervised ML algorithm that is fundamentally a type of artificial neural network (neural net) that belongs in the class of deep learning (Li et al., 2023). Along with Boltzmann machines and convolutional neural networks, autoencoders are one of the main types of models or algorithms used in deep learning (Li et al., 2023). The function of all neural nets is inspired by biological neural networks (Amari, 1972; Hopfield, 1982). Neural nets are variably connected layers of artificial neurons (nodes) that consist of input and output layers, hidden layers, neuron connections, connection weights and activation functions. Data are passed along connections between neurons, multiplied by weights, such that the weights for all inputs sum to unity at each neuron. The activation function defines neuronal output based on the sum of its inputs. The power of neural nets mainly stems from their ability to universally approximate functions. However, their capability depends on the number of neurons in the hidden layers. The layers of neurons perform linear and nonlinear transformations on data, whose purpose is to increasingly extract features through feature selection and extraction. Automated feature extraction functionally differentiates shallow learning and deep learning, whereby the former requires usually in-discipline knowledge to craft effective features, whereas the latter automates this process. Relative to other types of neural nets, autoencoders specialize in learning compact representations of data. A typical autoencoder architecture consists of two sections, the front end that faces the input data are the ‘encoder’ and the back end that regenerates the data is the ‘decoder’ (Fig. 4). The function of the encoder is to compress and encode data, by extracting increasingly compact features from original data. Hence, through training, patterns found in the original data are progressively captured in a compact latent space. The decoder reverses this process and reconstructs the input data to the extent possible. The error between the input data and the reconstructed data is measured using a metric, which

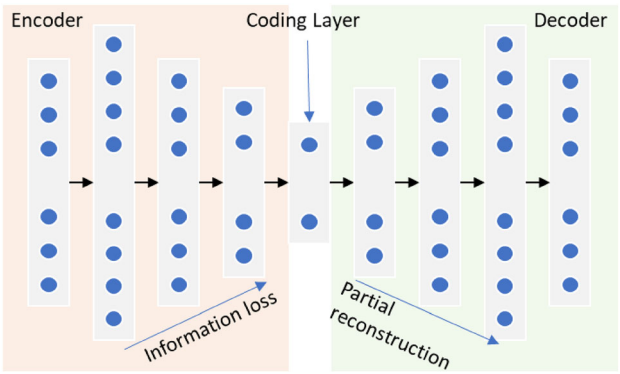


Figure 4. Network architecture of an autoencoder, featuring the encoder, coding layer and decoder sections. Through progressive lossy encoding, information is lost and only the most salient information is reconstructed.

is used to minimize the reconstruction error. The coding layer of the autoencoder is the 'bottleneck' and often the innermost neuronal layer between the front and back ends. Because the coding layer is the most compact layer in an autoencoder, its size is the constraint on the autoencoder in terms of its ability to model data. It determines whether an autoencoder is undercomplete (number of nodes in the coding layer smaller than original data dimension) or overcomplete (number of nodes in the coding layer exceeding original data dimension) (Bengio, 2009).

Autoencoders are used for a range of ML or AI tasks. Their ability to represent data in compact forms, without prior mathematical assumptions means that they could be used as a general-purpose dimensionality reduction tool (Li et al., 2023). In fact, one of the original uses of an autoencoder was as a general-purpose replacement of principal components analysis (Kramer, 1991). It is considered more general purpose than principal components analysis, mainly because it does not make model-specific (parametric) assumptions and as a type of neural net, it is theoretically capable of universal function approximation. The use of autoencoders in geochemistry is recent. Their current key application in geochemistry is primarily for the detection of multivariate geochemical anomalies as part of prospectivity and exploration (e.g., Zhang et al., 2019; Luo et al., 2020; Esmaeiloghli et al., 2023). Data reconstruction is used for anomaly detection, and the difference between reconstructed and actual data is a proxy for rare events in data, which depending on the context, could be class (3) or physical anomalies captured by geochemical data. By learning the most salient phenomena in the data, autoencoders preferentially replicate the most frequently observed characteristics. Consequently, the least observed characteristics are the worst reconstructed, which implies that by controlling an autoencoder's performance (or capability), it is possible to reconstruct data containing different levels of variability.

In the case of high-quality survey data, the largest source of variability is designed to be attributable to the variability of the sampled system (e.g., within class (3)). All other sources of variability, including data uncertainty (class (2)) should be smaller in comparison, because geochemical data are typically quality-controlled and assured during its generation to specifically maximize the amount of class (3) variability relative to that of class (2). Therefore, high-quality geochemical data should feature noise that is not generally spatially contiguous and of large amplitude relative to geochemical hotspots. Where the quality of survey data is poor, the QA/QC process should reveal that such data are precarious for mapping, for example, through ANOVA (Garrett, 1983; McCurdy & Garrett, 2016). Consequently, it is theoretically possible to reconstruct geochemical data, maximizing the preservation of class (3) variability, while minimizing class (2) variability. Therefore, denoising geochemical data can be formulated as a data reconstruction task using autoencoders, similar to denoising tasks in other domains because denoising is a key application of autoencoders (Vincent et al., 2010). However, this task is not solely an AI task, because the delineation between class (2) and class (3) variability is not inferable from the data alone and denoising is an ill-

posed inverse problem, which does not feature unique solutions. Therefore, there must be additional discipline-specific stopping criterion to guide the extent of reconstruction or information loss. The identification of a stopping criterion is also a key challenge in image denoising, because there are no objective methods to unambiguously assess and rank results (Fan et al., 2019).

Geochemical denoising can be formulated into a data reconstruction task with a geoscientific stopping criterion that is the result of a constraint imposed by the maximization of the loss of class (2) variability and the preservation of class (3) variability. For the data reconstruction task, each chemical analysis (e.g., nickel, copper or rare earth elements [REEs]) is fed into one of the input nodes, which becomes progressively encoded and transformed, such that at the coding layer, they become extracted or latent (geochemical) features (Fig. 4). Although unlike principal components analysis, the latent features are not geochemically interpretable, because they are not generally linear combinations, and therefore, do not correspond to the mineral and compositional variability of physical samples. However, this is not a concern, because the latent features are not used directly in a reconstruction task, and instead, they are decoded into data that are as identical to the input as possible, which are retrieved at the autoencoder's output. For the stopping criterion, it is possible to relate changes in reconstructed data (e.g., in the spatial domain–the nugget, or in the variable domain–autoencoder's performance) as a function of the autoencoder's coding layer size, to quality scores of the data. This is based on the reasoning that with increasing information loss through data reconstruction, there should be a transition from a loss of class (2) variability to class (3). Consequently, there may be qualitative behavioral changes (e.g., regime changes) in the extent of correlation between the original data quality (e.g., accuracy or precision) and the autoencoder's performance over a range of coding layer sizes.

An autoencoder contains several key hyperparameters, which capture mainly the architecture of the autoencoder—the number of hidden layers and the number of nodes in each layer. The front and back ends of autoencoders are not necessarily symmetric depending on use and heuristics of the architectural design. In addition, neuron activation function (activation) and the strength of regularization (α) are key hyperparameters. The activation could be a range of linear to sigmoidal functions that include the hyperbolic tangent (tanh), logistic and rectified linear (relu) functions. Regularization adds a penalty to the loss function by adding a squared sum of the weights of the network, multiplied by α. Another parameter is the ability to stop the training of the neural net early, if its performance fails to improve. This, along with regularization, reduces the tendency of the neural net to overfit. Overfitting is particularly an issue if the amount of training data is substantially less than the number of nodes in a neural net. A rule of thumb is that there should be 30 times the number of data points as compared to the number of nodes available (Amari et al., 1995). If this condition is unmet, early stopping of training increases the performance of the neural net by reducing overfitting. The initialization of most types of neural nets requires a randomly generated set of weights for all connections. Hence, the random number generator seed is a hyperparameter, and it could be varied to produce an ensemble of results, although its specific value is unimportant. The hyperparameter grid for this study is provided in Table 2. There is a variety of metrics that could be used to profile the autoencoder's performance, by comparing the difference between the input data and the predicted data. In this study, we utilized the coefficient of determination (CoD) or R^2, mean absolute percentage error (MAPE) and median absolute error (MedAE). By default, the MAPE metric is computed in fractions and not percentages. Detailed information regarding these metrics can be found at https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics. The computation workflow was implemented in Python and made use of ML algorithms and associated functions primarily in the Scikit-Learn package (Pedregosa et al., 2011).

RESULTS AND INTERPRETATIONS

An exhaustive grid search shows that the best performing autoencoder for the data has the architecture (56, 193, 67, 33, 29, 33, 67, 193, 56), with the tanh activation and early stopping enabled. In this configuration and averaging the results of 100 runs using different random initializations of the weights, the CoD score of the reconstruction (of all elements) was 0.941 with standard deviation of 0.003 (Fig. 5). However, the best performing autoencoder was generally not the most useful, because it reconstructed data noise, in addition to scientifically

Table 2. Parameter grid for the autoencoder used in this study

Parameter	Range
Network architecture	(56, 0 to 560, 56 to 112, 5 to 29, 56 to 112, 0 to 560, 56)
Activation	Relu, tanh, logistic
α	0.001, 0.0001, 0.00001
Early stopping	True, False

interesting events in data (e.g., mineral deposits). Keeping the configuration of all the other hidden layers fixed and lowering the number of nodes in the coding layer resulted in a reduction in the CoD score. Because variable domain metrics do not reflect the change in the structure of the data in spatial domain and geochemical survey data are generally intended to be used for mapping purposes (e.g., Govett, 1983; Demetrides et al., 2018), we also utilized metrics in the spatial domain. For this purpose, we computed the variograms of all elements (e.g., Fig. 6) and subsequently, computed the 'pseudo-nugget' effect. We defined the pseudo-nugget effect as the lowest non-zero value of the variogram at the shortest lag distance possible. This avoids making comparisons of the actual nugget effect, which requires a parametric model (e.g., spherical or Gaussian) to be fitted to the data. In this manner, the pseudo-nugget effect is a close approximation of the actual nugget effect in that both measure the spatial correlation of data at short distances, but no additional assumption of parametric approximation is required. Using the pseudo-nugget effect as a metric, we computed the correlation between the nugget of the reconstructed (predicted) data (Fig. 7). In particular, because the linear model that was fitted to the data showed a slope that was less than unity, there was a reduction in the pseudo-nugget effect (the model's slope is hereon referred to as the 'nugget slope' for brevity). Unfortunately, the sill could not be studied in the same manner, because it is generally dependent on range (e.g., cyclical structures), and it was not as uniquely defined as the pseudo-nugget effect. To obtain a statistically robust result and to also explore the control of the autoencoder's hidden layer size on the nugget effect, we performed 100 reconstructions of the data for each interval in a range of hidden layer sizes (from 29 to 5 nodes in intervals of 1 node). The results, presented in Figure 8, show that there was a definitive control of the hidden layer size on the performance of the autoencoder, as measured in the variable domain, which directly and linearly translates into a reduction in the nugget effect. In particular, it was observable that the relationship between the nugget slope and the autoencoder's CoD score was statistically robust (CoD = 0.97) and that the relationship was linear.

The reduction in the pseudo-nugget, in general, is indicative of a removal of noise because the spatial correlation was strengthened at smaller distances between samples. Consequently, it is expected that maps that are made using the denoised data would be more realistic, and especially, the short-ranged patterns observed in such maps would be more coherent and reliable. This is likely to be the case for geochemical anomalies or hotspots (e.g., mineral deposits), which should be much smaller in size compared to survey areas (Govett, 1983; Friske & Hornbrook, 1991; Friske, 1991). This would also be the case for per-block realism of resource-scale block models, because blocks are essentially small-scale features. However, this is difficult to assess qualitatively in the visual domain. The most optimal for visualization is probably cokriging, given our multivariate dataset (Tolosana-Delgado et al., 2011, 2019; Tolosana-Delgado & Boogaart, 2014). However, this complicates visual interpretations by necessitating also the interpretation of elemental correlations into otherwise univariate maps. Instead, we compared the maps of the reconstructed or predicted data with that of the original data using both inverse distance weighting and kriging. The purpose of inverse distance weighting is to establish that differences observed visually are attributable to changes in data, rather than changes in kriging parameters, such as the tuning of variograms and choice of kriging neighborhood parameters. Inverse distance weighting only requires a single hyperparameter, which is the power of the distance metric (set at 2 in this study). Kriging-based maps were also generated to provide the readers with a more realistic visualization of the data as it would be used for mapping purposes.

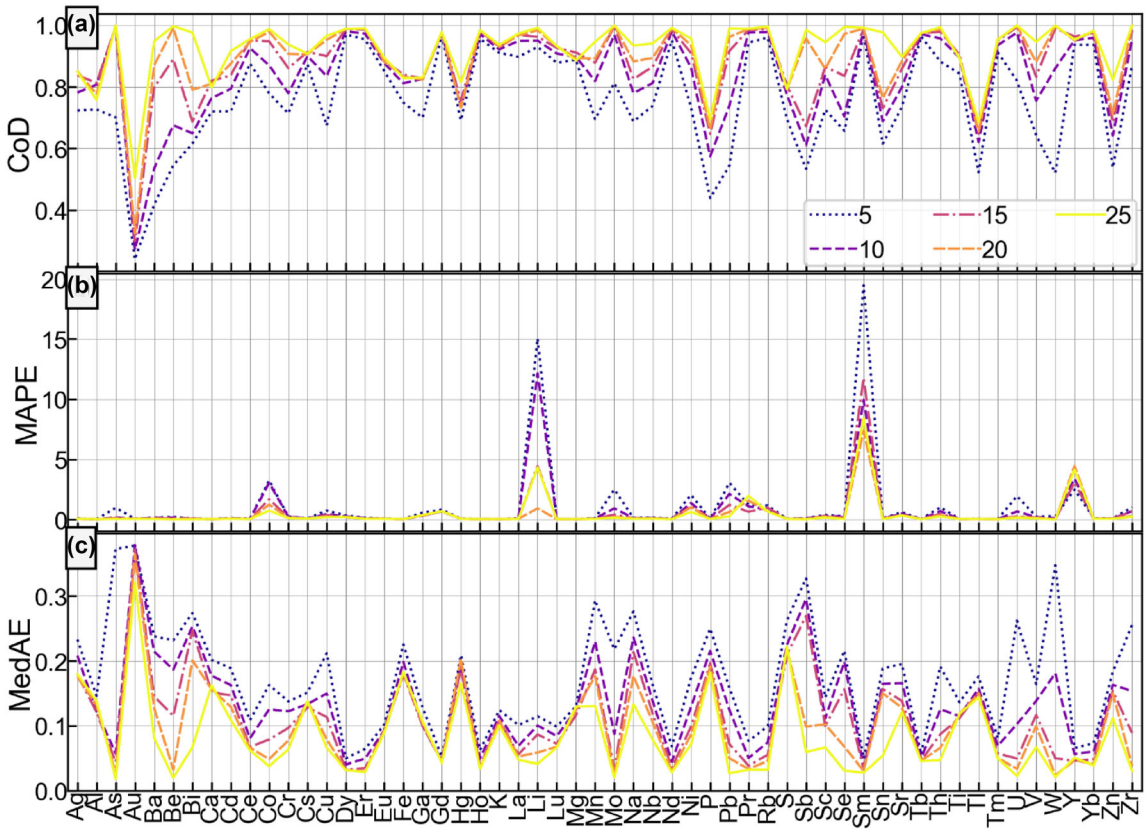


Figure 5. Autoencoder performance for coding layer sizes from 5 to 25 in intervals of 5, as assessed using the (**a**) CoD, (**b**) MAPE, and (**c**) MedAE metrics.

For the more rigorous visual comparison (Fig. 9), we chose a geochemical dataset that was re-analyzed in 2016 and included NTS zones 014D, 013 M and L, 023I and J (McCurdy et al., 2016; Fig. 2c). Notably, the re-analyzed region contains known glacial dispersal trains or eskers (approximately N70E) emanating from the REEs, Zr, Y, ± Be, Nb, U and Th enriched Strange Lake deposit (Geological Survey of Canada, 1980; Batterson et al., 1989; Zajac, 2015) as well as the southern portion of the Mistastin Batholith (Hammouche et al., 2012; Zhang et al., 2021a). Variances caused by other effects in the region (i.e., geomorphology, climate) are minimal as all samples are contained in the Canadian Shield and within similar climatic zones. Moreover, an exact degree of similarity between sediments is not known, but the lake sediment surveys were originally designed to cover the Canadian Shield area and intended for integrated use (Friske, 1991; Bourdeau & Dyer, 2023). The northern dispersal train or esker is visible for a variety of elements, including Be (Fig. 9a). A second dispersal train or esker to the south contains a different set of elements (Hammouche et al., 2012), but aside from its originating hotspot, it is not clearly defined on the Be map as compared to the one emanating from the Strange Lake deposit (Fig. 9a). Examining the changes induced by progressive decreases in the autoencoder's performance, as modulated through the coding layer size, it is obvious that the northern dispersal train or esker and some hotspots gradually weaken in intensity, becoming increasingly noticeable at coding layer sizes 11 and below (e.g., Fig. 9c, d and e). However, the high-intensity hotspot at the southern tip of the data coverage does not appear to fade with decreases in coding layer size. These observations indicate that the southmost hotspot occurs more frequently than the dispersal trains or eskers. Consequently, it is preferentially preserved with decreases in the autoencoder's coding layer size. It is also noticeable that the autoencoder removed a substantial amount

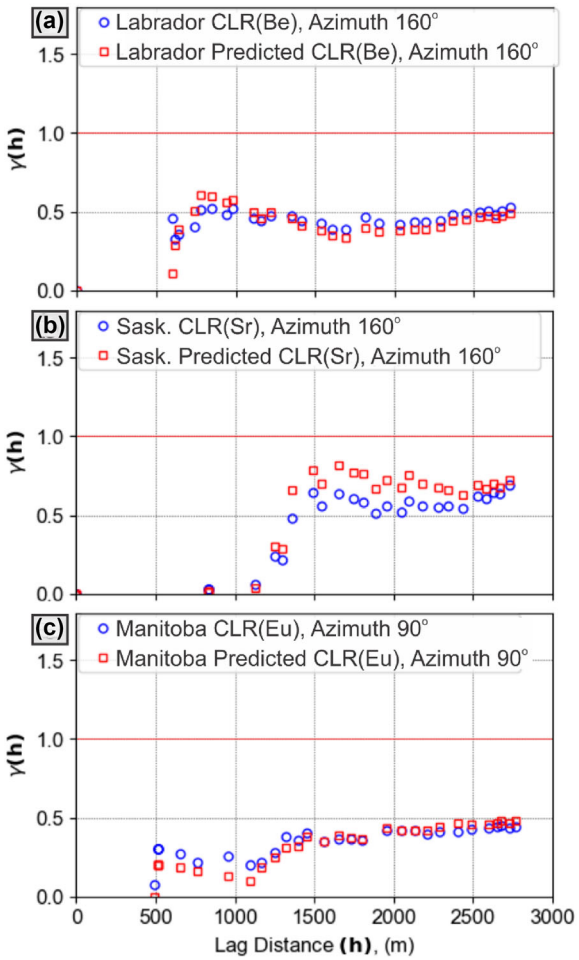


Figure 6. Variograms of (**a**) Be in Labrador, (**b**) Sr in Saskatchewan (Sask), and (**c**) Eu in Manitoba.

of background variability of the data, seen as lower chemical concentrations that undulate across the map (Fig. 9a). This effect is prominent at lower coding layer sizes (e.g., Figs. 9c, d and e). Effectively, this data reconstruction removed a substantial amount of the geochemical baseline, which consisted of noise and physically uninteresting events (e.g., actual anomalies due to random effects), and is the scientific basis for data reconstruction-based anomaly detection (e.g., Xiong and Zuo, 2016; Zhang et al., 2019; Luo et al., 2020). It is important to notice that, because the autoencoder trains and makes inferences solely in the variable domain, autoencoder-based denoising is characteristically different than spatial domain methods. For example, the

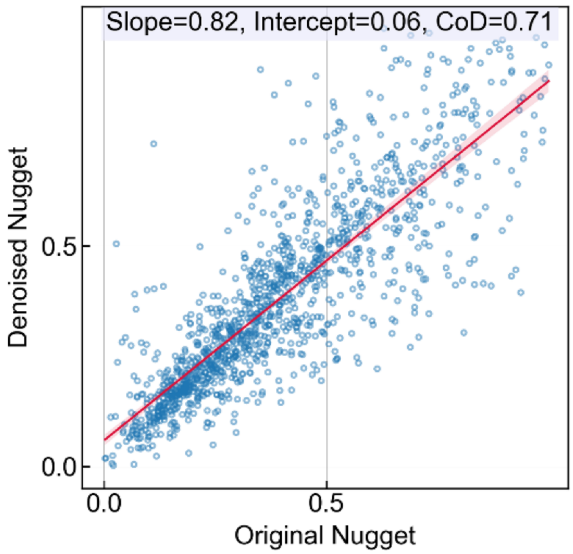


Figure 7. Comparison of the denoised nugget effect (the pseudo-nugget effect calculated using the reconstructed data) and the original (pseudo-) nugget effect.

autoencoder's results, when interpolated using an identical (controlled) method, do not exhibit blurring or softening of geochemical hotspots and treats every element differently. This is corroborated in the kriged maps (Figs. 10, 11 and 12), which suggest that denoised geochemical data are highly suitable for traditional mapping purposes. Image softening can occur with image post-processing, such as through the application of spatial filtering to gridded geochemical maps to remove low-frequency noise. This could impact downstream uses, for example, detecting linear dispersal patterns (Harris & Bonham-Carter, 2003). Subsequently, we computed the CoD, MAPE and MedAE metrics of the resulting maps as a function of the coding layer size (Fig. 13). It is clear that, in general, there was a loss in performance with decreasing coding layer size for many elements, and elements such as U, Ga and As were consistently the most affected. There was better correlation between the MedAE and CoD metrics as compared with that of MAPE and any other metric.

The relationship between data quality and autoencoder performance is important to understand for two reasons: (1) the identification of a robust stopping criterion, which sets a limit on the minimum number of nodes in the autoencoder's coding layer; and (2) to recognize the value of the denoised data and its usefulness for mapping. This

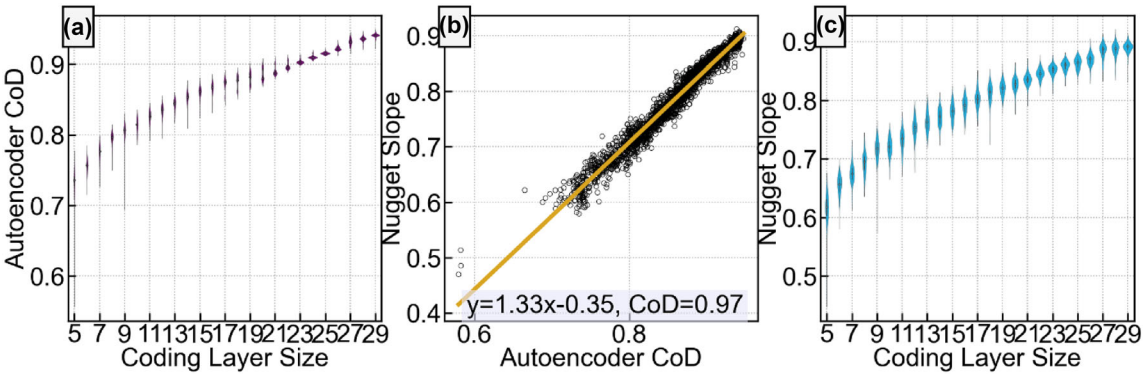


Figure 8. Effect of (**a**) the coding layer size on the autoencoder's performance as measured by the CoD metric, (**b**) the relationship between the nugget slope and the autoencoder's performance, and (**c**) the relationship between the nugget slope and the coding layer size.

can be analyzed by examining the correlation between the autoencoder's performance and quality measures of the data–accuracy, precision and the F-statistic in ANOVA. This information was provided as part of the QA/QC and data release packages (Table 1; McCurdy & Garrett, 2016). For this purpose, we used the statistically averaged results over 100 runs at various coding layer sizes and examined the relationship between the autoencoder's data reconstruction performance and the quality metrics on a per-element basis (for coding layer size at 29, see Fig. 14). It is clear that the older datasets in the Manitoba region from 2010 and 2013 (Table 1) are less accurate and precise than other datasets (Figs. 14a, b). Consequently, the datasets did not exhibit comparable data uncertainty. Moreover, in general, there was a negative relationship between the MedAE score and the accuracy of the input data (Fig. 14a). The relationship between the MedAE scores and the precision and F-statistics of the ANOVA was less obvious (Figs. 14b, c). The relationships between the other performance metrics and the data quality metrics were similar. Because the autoencoder's behavior is an emergent property of the neural net, it is not possible to understand the exact mechanisms of the network in a causal or reductive sense. Consequently, we made no assumptions of the relationship between the autoencoder's performance and the data quality metrics, and instead, quantitively studied the relationships using the Spearman rank correlation, excluding the Manitoba 2010 datasets, because of its poor comparability with other datasets (Fig. 15). It is clear that there was a systematic and consistent relationship across the data quality metrics examined. In particular, there were several qualitative behavioral changes in the relationships: (1) at 19 nodes; and (2) between 12 to 11 nodes in the coding layer as seen in Figure 15. These were regime changes in the behavior of the autoencoder because the dependence of the autoencoder's performance and the data quality changed sharply and consistently. Type (1) change in particular, transitioned from a steeper slope before 19 nodes in the coding layer to a shallower slope at fewer nodes. This behavior was the most prominent for the MAPE metric (Fig. 15b), but was also noticeable for the CoD and MedAE metrics (Figs. 15a, c). As the autoencoder's coding layer size continued to decrease past 19 nodes, the effect of data uncertainty associated with spatial variability, accuracy and precision saturated. Consequently, below 19 nodes, the autoencoder was progressively removing other types of anomalies aside from uncertainty. Type (2) change was primarily associated with the ANOVA's F-statistic. In particular, type (2) change was not obvious or as significant for either accuracy or precision. This implies that the uncertainty in the spatial variability of the data was maximally removed at a coding layer size of 11. Therefore, at coding layer sizes below 11 nodes, the removal of anomalies in the data changed from primarily removal of data uncertainty to removal of physically relevant anomalies. Therefore, a coding layer size of 11 is a stopping criterion that separates the autoencoder's functionality from primarily removing class (2) variability to class (3) variability.

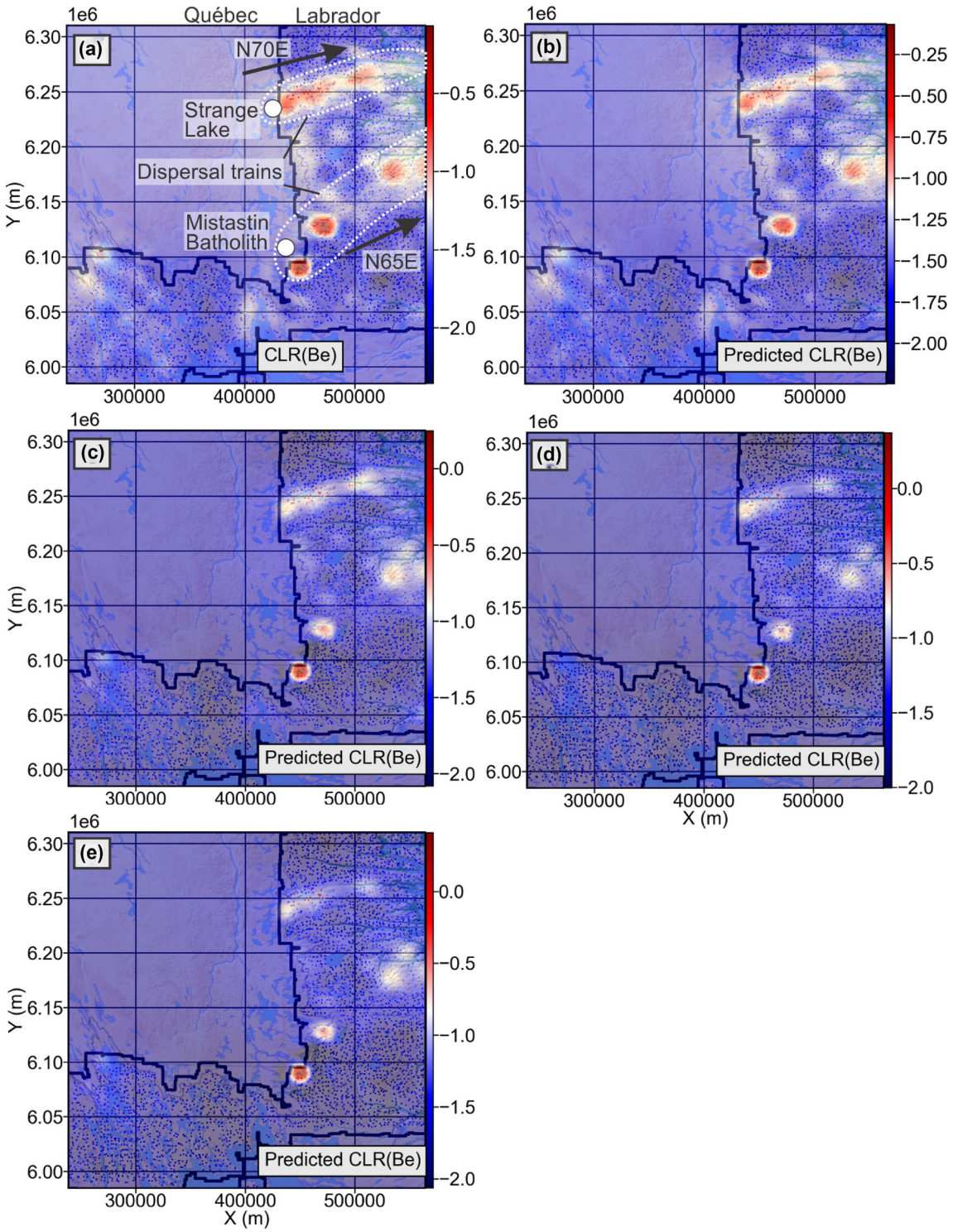


Figure 9. Maps of the original data for Be in the 2016 re-analyzed dataset in Labrador (Fig. 2a) in (**a**), average of predicted (or reconstructed) data using hidden layer sizes of 16 in (**b**), 13 in (**c**), 9 in (**d**) and 5 in (**e**). The black arrow in (**a**) corresponds to the direction of glacial dispersal trains or eskers.

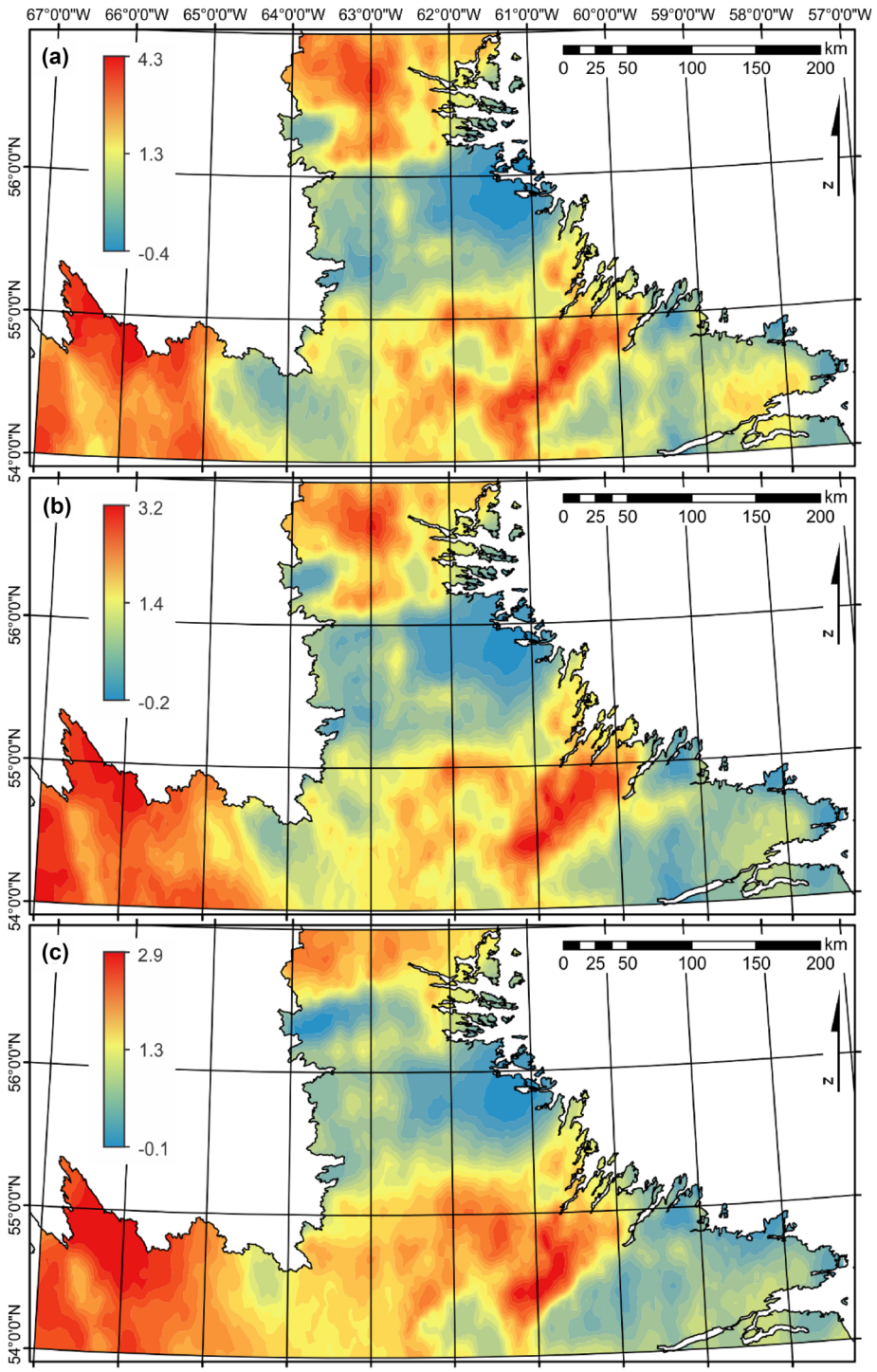


Figure 10. Spatial interpolation maps of copper using the original Labrador dataset (Fig. 2c) in (**a**). Average of predicted (or reconstructed) data using hidden layer sizes of 13 in (**b**) and 5 in (**c**). The interpolation method used here was simple kriging. Hyperparameters for (**a**) were nugget = 0.59, sill = 0.45, range = 1.48 km, using the stable model (similar for (**b**) and (**c**)). Values correspond to CLR-transformed ppm.

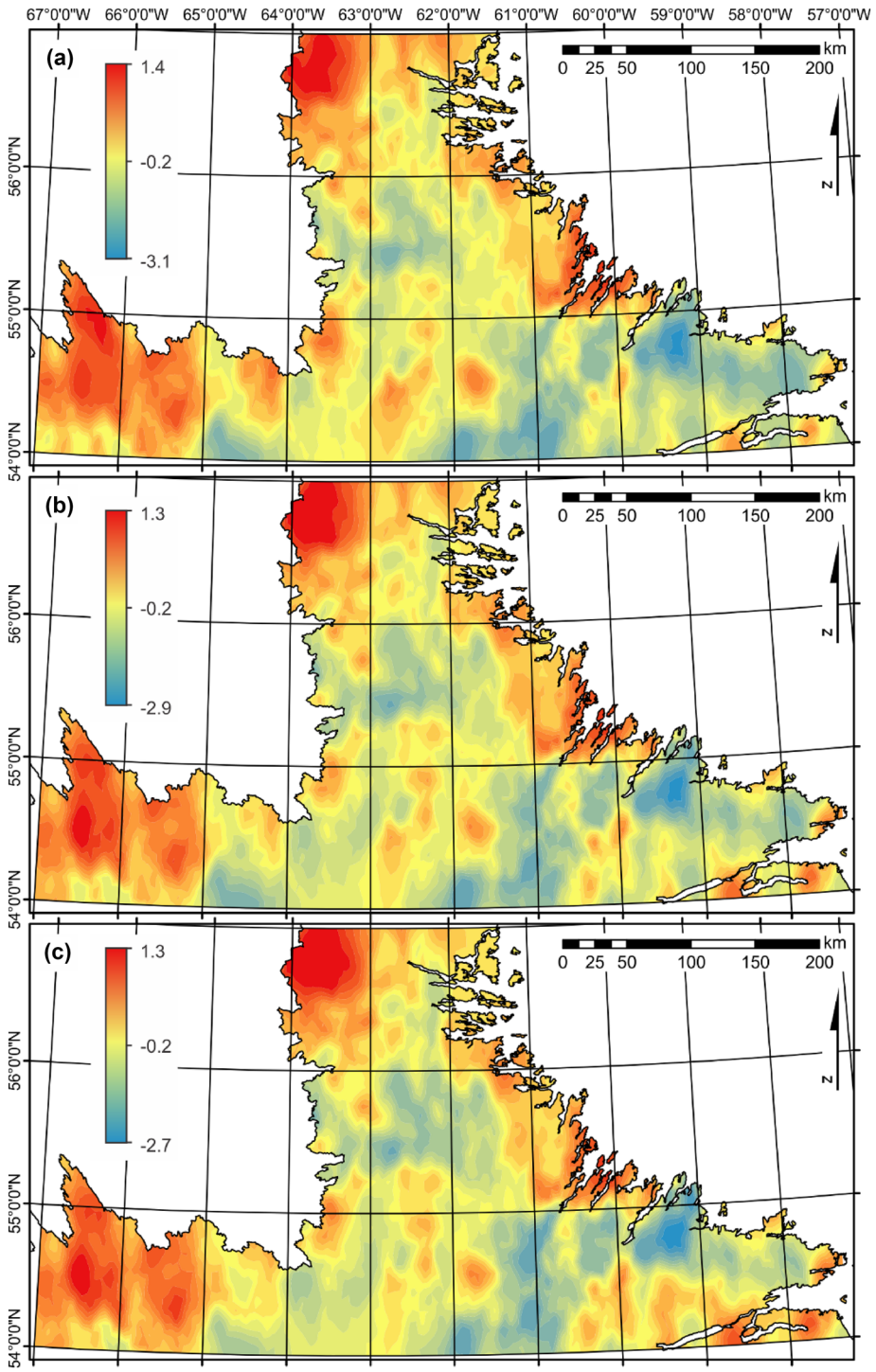


Figure 11. Spatial interpolation maps of lithium using the original Labrador dataset (Fig. 2c) in (**a**). Average of predicted (or reconstructed) data using hidden layer sizes of 13 in (**b**) and 5 in (**c**). The interpolation method used here was simple kriging. Hyperparameters for (**a**) were nugget = 0.64, sill = 0.25, range = 0.97 km, using the stable model (similar for b and c). Values correspond to CLR-transformed ppm.

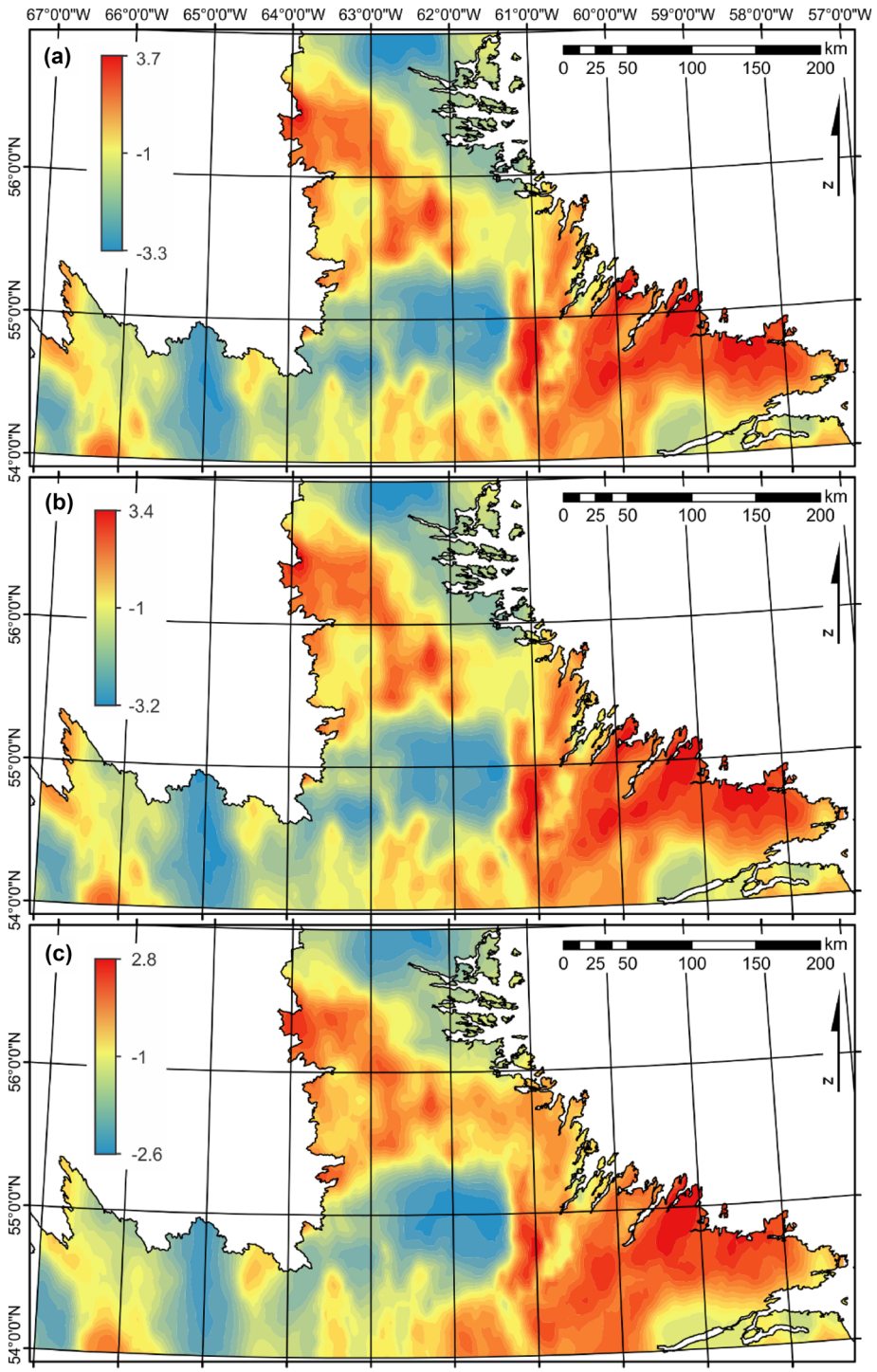


Figure 12. Spatial interpolation maps of uranium using the original Labrador dataset (Fig. 2c) in (**a**). Average of predicted (or reconstructed) data using hidden layer sizes of 13 in (**b**) and 5 in (**c**). The interpolation method used here was simple kriging. Hyperparameters for (**a**) were nugget = 0.35, sill = 0.66, range = 1.40 km, using the stable model (similar for (**b**) and (**c**)). Values correspond to CLR-transformed ppm.

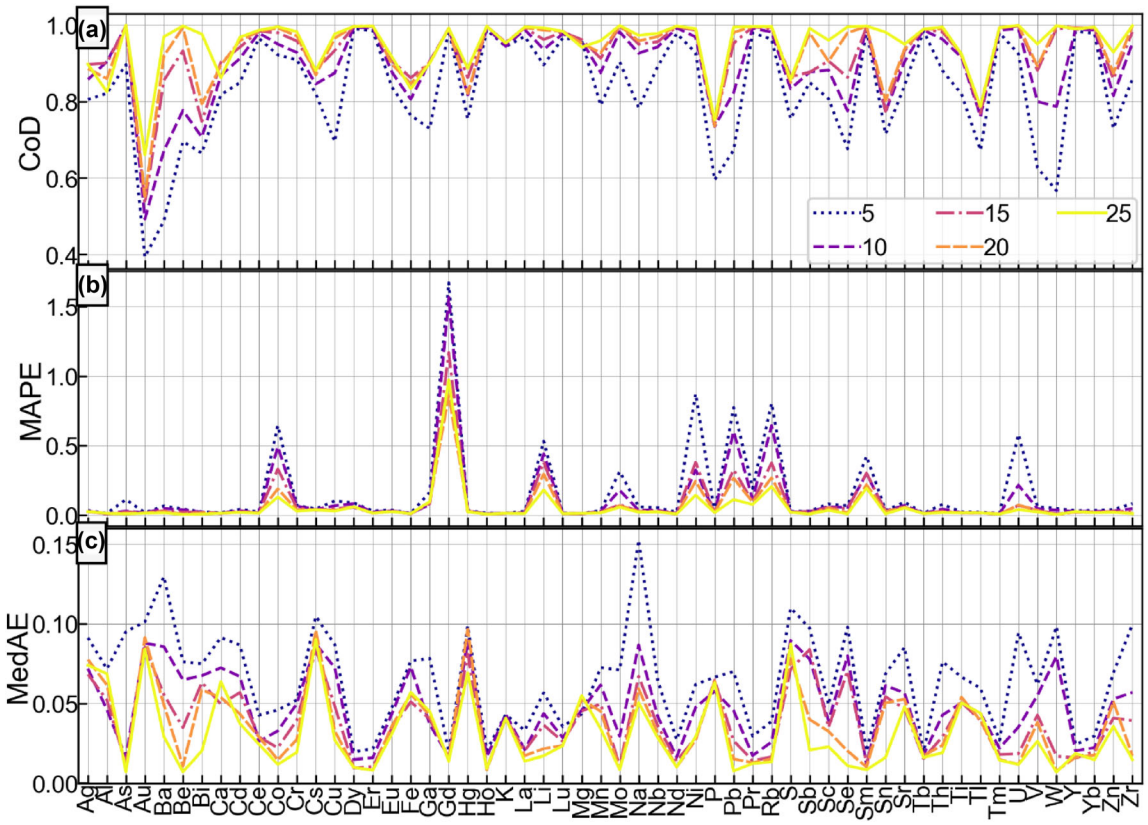


Figure 13. Image domain metric scores for the 2016 re-analyzed dataset in Labrador, for each element as a function of the coding layer size (from 5 to 25 in intervals of 5 as assessed using the (**a**) CoD, (**b**) MAPE, and (**c**) MedAE metrics.

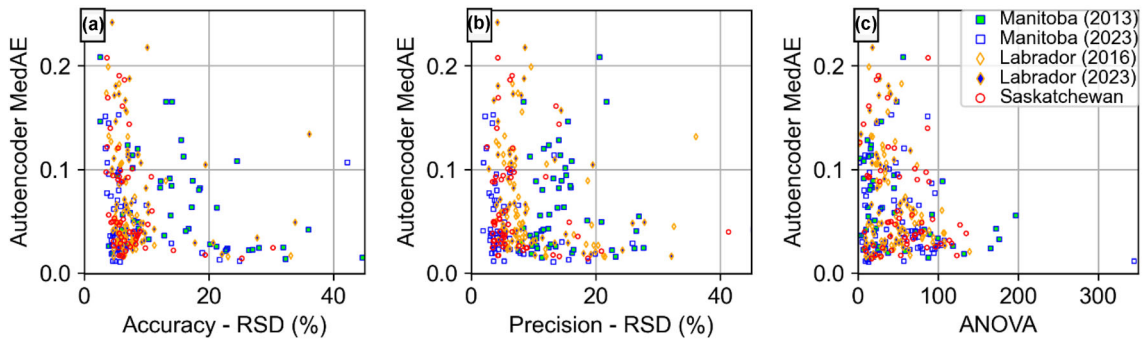


Figure 14. Performance of the autoencoder as measured using the MedAE metric versus (**a**) the accuracy of the input data, (**b**) precision of the input data, and the (c) F-statistic of the ANOVA of the input data. The label Manitoba 2013 also includes the dataset from 2010.

DISCUSSION

Uncertainty in Geochemical Data

Based on the results and our analysis, there are several key findings: (1) autoencoders can be used to reduce or modulate the uncertainty in the data; (2) reduction in uncertainty is observable in the spatial domain as a decrease of the nugget effect; and (3) it is possible to infer the behavioral regime of the autoencoder between the removal of data uncertainty associated with data quality metrics and the

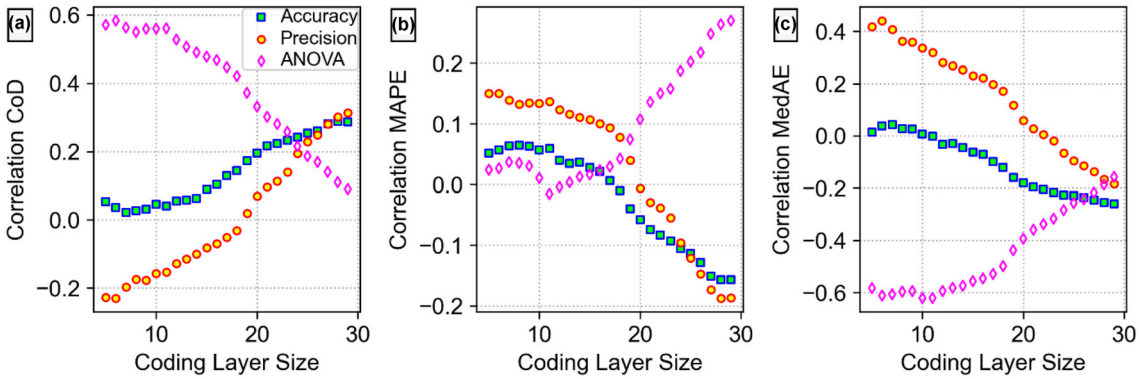


Figure 15. Spearman rank correlation coefficient between the autoencoder performance as assessed using the (**a**) CoD, (**b**) MAPE and (**c**) MedAE metrics, and data quality metrics versus the coding layer size.

removal of relevant events in data, such as meaningful anomalies. Our analysis of the autoencoder results demonstrates an identifiable and sharp transition between the removal of uncertainty in data and removal of larger data variability. It is possible to use this information to infer an upper bound limit to the amount of data uncertainty that exists in the data. Below 19 coding layer nodes, the data reconstruction showed evidence of saturated sensitivity with respect to the quality of the data, which implies that roughly 88% of the variability of the data were unrelated to data noise or uncertainty (Fig. 8a). This implies that the nugget slope that could be achieved using denoising was about 0.82, which corresponded to a 18% reduction in the pseudo-nugget effect. This extent of reduction could have significant impacts on downstream uses of such data for mapping purposes, because a decreased nugget effect decreases the level of smoothing of maps (or resource models) thus increasing the local accuracy of such products. It is important to note that our method to infer the transition from the removal of class (2) to class (3) variability is presumably broadly applicable in geochemistry, provided that the quality of data was originally assessed at the time of its generation. Consequently, it would be possible in general to apply our method to reduce geochemical data noise while minimizing the removal of potentially physically relevant events captured by such data.

The effect of denoising is unevenly distributed across the elements, which is expected given that the autoencoder's performance is correlated with the accuracy, precision and the F-statistic of the input data. Hence, for elements that exhibited the lowest quality, there was a generally larger extent of denoising by the autoencoder. Therefore, data reconstruction performance is an indicator of post-hoc data quality. In this sense, it is possible to use our method to detect data quality even without primary QA/QC information. In our case, the autoencoder's reconstruction performance was poor for a number of elements (Fig. 5). Interpreting this observation in the data-driven domain is possible, because reconstruction performance for multivariate tasks is proportional to the robustness of high-dimensional feature correlations. Consequently, concentrations of elements that were constructed poorly were not as strongly related to the concentrations of other elements in a nonlinear and high-dimensional sense. Because autoencoders are generalizations of principal components analysis, the nonlinear and high-dimensional relationships are generalizations of linear elemental associations. There are two major causes at the data level for observed differences in reconstruction performance by considering generalized multivariate relationships: (1) relationships are inconsistent; and (2) relationships are rare. Inconsistent relationships can arise from natural or sampling-induced processes, while rare relationships appear as isolated occurrences that are difficult to distinguish from noise.

A thorough geoscientific investigation as to the autoencoder's reconstruction performance for specific elements (e.g., elemental associations, sample digestion methods, bedrock geology) remains to be conducted, which is primarily knowledge-based and outside of the domain of a data-driven methodology study. Here, we merely highlight some

potential natural causal mechanisms for poor performance for the worst five elements according to both CoD and MedAE metrics, which are: Au, Ba, P, Sb and W (Figs. 5a, c). Elements Au and W are well known to occur as coarse grains or nuggets, leading to a heterogeneous distribution within any surficial media (Dyer & Barnett, 2007; Dominy, 2014). This can give rise to a high-nugget effect, because the underlying prior is discontinuous. This type of noise cannot be generally removed without spatial averaging, because repeated sampling at the same location does not remove the nugget effect. Elements Ba and P are known to be highly mobile in weathering environments (Middelburg et al., 1988). Consequently, they can give rise to differential multivariate relationships, for example, those occurring in weathered versus in non-weathered environments, which results in covariate shift. Similarly, Sb has a strong affinity to sulfur, and thus, considered a chalcophile element as per Goldschmidt's classification (Goldschmidt, 1937). Figure 5a and c demonstrate that all determined chalcophile elements (i.e., Ag, As, Bi, Cd, Cu, Ga, Hg, Pb, Sb, Se, Sn, Tl and Zn) yielded poorer reconstruction performance. Chalcophile elements are dominantly found in sulfide group minerals, which themselves are known to readily and quickly weather, rendering chalcophile elements highly mobile in weathering environments. Differential weathering by location and age of sediment is a possibility that introduces data variance. In general, all things being equal (e.g., amount of data), increased data variability means that regardless of the model, reconstruction would be worse. This can be countered explicitly by recognizing the potential for this type of variability and designing surveys to increase data coverage or coverage rate in regions suspected of exhibiting more geological or geochemical diversity. However, controlling this type of data bias is a very recent concern and certainly was not within the purview of the design of the surveys. Conversely, data quality is substantially better for some elements, notably Dy, Er, Li, Nd and Sm. All of these aforementioned elements are known to be highly immobile during weathering as these elements are found primarily within resistive mineral phases (e.g., zircon, monazite, niobate) (Crock and Lamothe, 2011). Consequently, their multivariate relationships are more stable over the study areas (a type of relationship stationarity).

Implications for Data-Driven Uses of Geochemical Data

All types of deep learning algorithms generally require nonlinear optimization, which is associated with a tendency of a model's parameters to settle into local minima (e.g., Hsieh, 2009). This fact, combined with the ability of neural nets to approximate any function arbitrarily well implies that invariably, noise such as data uncertainty can be captured easily by trained models, which results in overfitting (Hsieh, 2009). This matter is made worse in the domain of geospatial data, where trained models may be expected to be deployed over spatial distances, a task known as geostatistical transferred learning. Notably, geostatistical transferred learning is unsolved because of various issues such as spatial covariate shifts (Hoffimann et al., 2021). In addition to covariate shifts, although geoscientific surveys typically are standardized as much as possible, uncontrollable changes in natural variability imply that batches of such data may also have covariate shifts in data quality, such as data precision and accuracy. Consequently, overfitting of models using geochemical data is probably impossible or very difficult to detect using standard ML approaches (e.g., cross-validation), where data contain incomparable uncertainty. This is known to be the case for the data used in this study, because the data quality clearly vary between and within Saskatchewan, Labrador (2016 and 2023) and Manitoba (2010, 2013 and 2023) datasets (e.g., Fig. 14). Shifts in data quality are an additional and higher order covariate shift that also complicates geostatistical transferred learning. In other domains, concerns around local minima and overfitting led to controlled noise being intentionally added to training data to aid model generalization for neural networks (e.g., Bishop, 1995; Reed & Marks II, 1999; Goodfellow et al., 2016) and increasingly to ensembles of neural nets (Yang et al., 2021). This approach could be feasible for standardizing geochemical data uncertainty, although it would require a rigorous model of the nature of multivariate geochemical noise, which would obviously differ with changes in protocol components, such as instruments and media types. However, this is not a very practical approach, because even if it were technically feasible, adding noise to data decreases the value of each data point, which is very costly in most of geosciences (e.g., Demetrides et al., 2018). With present practices, it is difficult to minimize overfitting for complex neural

nets trained on geochemical surveys because, even at the national scale, the amount of data that were available for this study was substantially less than 30 times the number of neurons in our model.

It is important to note that autoencoders and other deep learning algorithms are all neural nets with the exception that autoencoders are, by design, intended to capture only salient events in data. Therefore, because autoencoders are purposefully designed to remove noise, which at the class (2) level includes variability caused by data uncertainty, the performance of an autoencoder in a data denoising task is a strong indicator of the level of impact of uncertainty in geochemical data on other types of neural nets. In this view, if for example, about 12% of the variability in the data is indeed irreducible uncertainty, then only overcomplete autoencoders, or other neural nets that are overfitted, can capture beyond 88% of the variability in the data. This would also affect traditional data analysis techniques. For example, for principal components-based data analysis or dimensional reduction tasks, the maximal amount of variability that one should strive to retain could be, in theory, determined through an autoencoder denoising analysis (e.g., 88% in the case of our dataset). This would provide an objective constraint to the delineation of geochemical background. Consequently, autoencoders may be desirable to denoise data before application of downstream data analysis and deep learning algorithms. This is because data denoising prior to its usage results in: (1) better generalizability of models downstream; (2) better leveling of data uncertainty across datasets; and (3) an ability to modulate data uncertainty to determine the uncertainty of downstream products through uncertainty propagation.

Without the autoencoder-based denoising process, stochastic methods can be used to inject structured noise into data to modulate outcomes (e.g., Zhu et al., 2020; Cao et al., 2022). However, because geochemical data are generally multivariate, its uncertainty is also multivariate. Univariate adjustment of elemental concentrations does not replicate the type of uncertainty typically associated with multivariate data, because individual adjustments are uncorrelated both in the variable and spatial domains. Denoising the data in variable domain clearly impacts spatial domain characteristics, because we observed that denoising the data in variable domain increased spatial correlation of the data at short ranges. The direct implication of this observation is that maps that are made using denoised data are likely to be more reliable because they are less noisy than maps made using the original data. This means that, for downstream applications such as mineral prospectivity mapping, models that are fitted to denoised data are more likely to capture salient features instead of noise. The impact of this is probably the most noticeable in applications where the training data features sparse labels, which can be the case for many types of mineral deposits. Removing unnecessary uncertainty in evidence layers prior to performing modeling may result in more realistic prospectivity maps, although this remains to be tested.

SUMMARY AND CONCLUSIONS

There are several key findings of this study. We determined that: (1) it is possible to remove noise in geochemical data without incurring interpolation, which permits a denoise-then-interpolate approach to using geochemical data; (2) autoencoders effectively denoise geochemical data up to a limit, which delineates the removal of primarily class (2) to (3) variability; and (3) it is possible to analyze performance regimes of an autoencoder, thereby providing an objective stopping criterion to determine the coding layer size. We showed that the denoised data featured improved spatial correlation at short ranges as evidenced by a general reduction in the nugget effect. Moreover, there is an empirical linear relationship between the reduction in nugget and the autoencoder performance. We determined that the autoencoder's performance is relatable to data quality metrics that include accuracy, precision and F-statistic (ANOVA) of the original data. A key implication is that, because uncertainty exists in geochemical data and presumably for all measured data, its effect should be studied for downstream uses. Data leveling, for example, may be extended to match beyond key statistical moments of data, as leveling is not yet considered for data uncertainty but has significant implications for model generalizability (e.g., training models on heteroscedastic errors) and therefore, the realism of downstream products employing complex models. This is especially important for modern downstream uses of such data because complex models can capture data uncertainty in addition to desirable events and the performance of geospatially transferred models are potentially significantly over-estimated due to the

unsolved geostatistical transferred learning problem. Our method is a simple, clear and consistent approach to removing geochemical data noise with the amount of denoising guided by scientific data quality. Variably denoised data, as well as the original data can be fed into a single downstream workflow (e.g., mapping) and the differences in the outcome can be quantified to propagate data uncertainty.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Christopher J.M. Lawley and Dr. Nikole Bingham-Koslowsky (Geological Survey of Canada) for discussions and comments which have improved the scope of this study. We would also like to thank the sedimentological laboratory and Stephen J.A. Day (Geological Survey of Canada) for their lead in the re-analysis of legacy lake sediment samples. The authors thank three anonymous reviewers for their constructive comments which have greatly improved this manuscript. We are also thankful to Prof. E.J.M. Carranza for editorial handling.

FUNDING

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

DECLARATIONS

Conflict of Interest The authors declare no known conflicts of interests that could appear to have influenced this work.

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

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall.

Amari, S. I. (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers, 100*(11), 1197–1206.

Amari, S. I., Murata, N., Müller, K. R., Finke, M., & Yang, H. (1995). Statistical theory of overtraining–Is cross-validation asymptotically effective? *Advances in Neural Information Processing Systems, 8*, 195.

AusIMM (2014). Mineral resource and ore reserve estimation: the AusIMM guide to good practice. Monograph 30. *The Australian Institute of Mining and Metallurgy* (2nd ed). Australia.

Balaram, V., & Subramanyam, K. S. V. (2022). Sample preparation for geochemical analysis: Strategies and significance. *Advances in Sample Preparation, 1*, 100010. https://doi.org/10.1016/j.sampre.2022.100010.

Batterson, M. J., DiLabio, R. N. W., & Coker, W. B. (1989). Glacial dispersal from the Strange Lake alkalic complex. Northern Labrador. *Drift Prospecting*, *89*(20), 31–40. https://doi.org/10.4095/127362.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1–127. https://doi.org/10.1561/2200000006.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University Press.

Bourdeau, J. E., & Dyer, R. D. (2023). Regional-scale lake-sediment sampling and analytical protocols with examples from the Geological Survey of Canada. *Geological Survey of Canada*, Open File 8980. https://doi.org/10.4095/331911.

Cao, J., Ma, J., Huang, D., Yu, P., Wang, J., & Zheng, K. (2022). Method to enhance deep learning fault diagnosis by generating adversarial samples. *Applied Soft Computing, 116*, 108385. https://doi.org/10.1016/j.asoc.2021.108385.

Carranza, E. J. M. (2011). Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. *Journal of Geochemical Exploration, 110*(2), 167–185.

Carrasco, P. (2010). Nugget effect, artificial or natural? *Journal of the Southern African Institute of Mining and Metallurgy, 110*(6), 299–305.

Clark, I. (2010). Statistics or geostatistics? Sampling error or nugget effect? *Journal of the Southern African Institute of Mining and Metallurgy, 110*(6), 307–312.

Crock, J. G., & Lamothe, P. J. (2011). Inorganic chemical analysis of environmental materials - a lecture series. *U.S. Geological Survey*, Open File 2011-1193. https://doi.org/10.3133/ofr20111193.

Davis, C. (2010). GPS-like navigation underground. In *IEEE/ION Position, Location and Navigation Symposium* (pp. 1108-1111). IEEE. https://doi.org/10.1109/PLANS.2010.5507196.

Demetrides, A., Smith, D. B., & Wang, X. (2018). General concepts of geochemical mapping at global, regional, and local

scales for mineral exploration and environmental purposes. *Geochimica Brasiliensis, 32*(2), 136–179.

Dominy, S. C. (2014). Predicting the unpredictable - evaluating high-nugget effect gold deposits. In *Monograph 30 - Mineral Resource and Ore Reserve Estimation - The AusIMM guide to good practice* (2nd ed., pp. 659–678). AusIMM.

Dyer, R., & Hamilton, S. (2007). The Ontario Geological Survey Lake sediment geochemical program: progress towards a geochemical map of Ontario, Canada. In *EXPLORE Newsletter* (vol. 135, pp. 2–11). Association of Applied Geochemists.

Dyer, R. D., & Barnett, P. J. (2007). Multimedia exploration strategies for PGEs: insights from the surficial geochemistry case studies project, Lake Nipigon region geoscience initiative, northwestern Ontario. *Canadian Journal of Earth Sciences, 44*(8), 136–179.

Esmaeiloghli, S., Tabatabaei, S. H., & Carranza, E. J. M. (2023). Infomax-based deep autoencoder network for recognition of multi-element geochemical anomalies linked to mineralization. *Computers & Geosciences, 175*, 105341.

Fan, L., Zhang, F., Fan, H., & Zhang, C. (2019). Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art, 2*, 1–12.

Friske, P. W. B. (1991). The application of lake sediment geochemistry in mineral exploration. In J. M. Franklin, J. M. Duke, W. W. Shilts, W. B. Coker, P. W. B. Friske, Y. T. Maurice, S. B. Ballantyne, C. E. Dunn, G. E. M. Hall & R. G. Garrett (Eds.), *Exploration Geochemistry Workshop* (pp. 157–180). Geological Survey of Canada, Open File 2390. https://doi.org/10.4095/132392.

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

Garrett, R. G. (1983). Sampling methodology. In R. J. Howarth (Ed.), *Statistics and data analysis in geochemical prospecting* (pp. 83–100). Elsevier.

Geboy, N. J. & Engle, M. A. (2011). Quality assurance and quality control for geochemical data: a primer for the research scientist. *U.S. Geological Survey*, Open File 2011-1187.

Geological Survey of Canada. (1980). Airborne Gamma ray spectrometric map, Dihourse Lake Quebec-Newfoundland. *Geological Survey of Canada, 35124*(08), 128. https://doi.org/10.4095/124881.

Goldschmidt, V. M. (1937). The principles of distribution of chemical elements in minerals and rocks. *Journal of the Chemical Society (Resumed), 37*, 655–673. https://doi.org/10.1039/JR9370000655.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Govett, G. J. S. (1983). Geochemistry in the exploration sequence. In G. J. S. Govett (Ed.), *Handbook of exploration geochemistry* (pp. 7–15). Elsevier.

Grunsky, E.C., & de Caritat, P. (2017). Advances in the use of geochemical data for mineral exploration. In V. Tschirhart, & M. D. Thomas (Eds.), *Proceedings of exploration 17: sixth decennial international conference on mineral exploration* (pp. 441–456). Decennial Minerals Exploration Conferences.

Grunsky, E. C. (2010). The interpretation of geochemical survey data. *Geochemistry Exploration Environment Analysis, 10*(1), 27–74.

Grunsky, E. C., & de Caritat, P. (2020). State-of-the-art analysis of geochemical data for mineral exploration. *Geochemistry Exploration Environment Analysis, 20*(2), 217–232.

Guan, Q., Ren, S., Chen, L., Feng, B., & Yao, Y. (2021). A spatial-compositional feature fusion convolutional autoencoder for multivariate geochemical anomaly recognition. *Computers & Geosciences, 156*, 104890.

Guartán, J. A., & Emery, X. (2021). Regionalized classification of geochemical data with filtering of measurement noises for predictive lithological mapping. *Natural Resources Research, 30*, 1033–1052.

Hammouche, H., Legouix, C., Goutier, J., & Dion, C. (2012). Géologie de la région du lac Zeni. *Ministère des Ressources naturelles, Québec*, RG 2012-02.

Harris, J. R., & Bonham-Carter, G. F. (2003). A method for detecting glacial dispersal trains in till geochemical data. *Geochemistry Exploration Environment Analysis, 3*(2), 133–155.

He, Y., Zhou, Y., Wen, T., Zhang, S., Huang, F., Zou, X., Ma, X., & Zhu, Y. (2022). A review of machine learning in geochemistry and cosmochemistry: method improvements and applications. *Applied Geochemistry, 140*, 105273.

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

Hofmann, T., Darsow, A., & Schafmeister, M. T. (2010). Importance of the nugget effect in variography on modeling zinc leaching from a contaminated site using simulated annealing. *Journal of Hydrology, 389*(1–2), 78–89.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*(8), 2554–2558.

Hsieh, W. W. (2009). *Machine learning methods in the environmental sciences: Neural networks and kernels*. Cambridge University Press.

Ilesanmi, A. E., & Ilesanmi, T. O. (2021). Methods for image denoising using convolutional neural network: A review. *Complex & Intelligent Systems, 7*(5), 2179–2198.

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

Knight, R. D., Kjarsgaard, B. A., & Russell, H. A. (2021). An analytical protocol for determining the elemental chemistry of quaternary sediments using a portable x-ray fluorescence spectrometer. *Applied Geochemistry, 131*, 105026.

Knott, J. M., Glysson, G. D., Malo, B. A., & Schroder, L. J. (1993). Quality assurance plan for the collection and processing of sediment data by the U.S. geological survey, water resources division. *US Geological Survey, 92*, 499. https://doi.org/10.3133/ofr92499.

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

Li, P., Pei, Y., & Li, J. (2023). A comprehensive survey on design and application of autoencoder in deep learning. *Applied Soft Computing, 138*, 110176.

Liu, B.L., Wang, X.Q., Guo, K., & Zhao, Y.H. (2014). Geochemical data processing based on wavelet de-noising. In *2014 11th international computer conference on wavelet actiev media technology and information processing (ICCWAMTIP)* (pp. 144-147). IEEE. https://doi.org/10.1109/ICCWAMTIP.2014.7073379.

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

McCurdy, M. W., Spirito, W. A., Grunsky, E. C., Day, S. J. A., McNeil, R. J., & Coker, W. B. (2014). The evolution of the geological survey of Canada's regional reconnaissance geochemical drainage sediment and water surveys. In *EXPLORE Newsletter*. Association of Applied Geochemists. pp. 1–10.

McCurdy, M. W., Amor, S. D., & Finch, C. (2016). Regional lake sediment and water geochemical data, western and central Labrador (NTS 13-L, 13-M, 14-D, 23-I and 23-J). *Geological Survey of Canada*. https://doi.org/10.4095/298834.

McCurdy, M. W., Campbell, J. E., Bohm, C. O., Trommenlen, M. S., & Syme, E. C. (2013). Regional lake sediment geochem-

ical data, Kasmere Lake-Nueltin Lake area, Manitoba, NTS 64-K, 64N, 64-O. *Geological Survey of Canada*. https://doi.org/10.4095/292267.

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

Middelburg, J. J., van der Weijden, C. H., & Woittiez, J. R. (1988). Chemical processes affecting the mobility of major, minor and trace elements during weathering of granitic rocks. *Chemical Geology, 68*(3–4), 253–273.

Nwaila, G. T., Zhang, S. E., Bourdeau, J. E., Frimmel, H. E., & Ghorbani, Y. (2023). Spatial interpolation using machine learning: from patterns and regularities to block models. *Natural Resources Research*. https://doi.org/10.1007/s11053-023-10280-7.

OSC, 2011. OSC Bulletin, v. 34, Chap. Rules and Policies: NI 43-101 standards of disclosure for mineral projects. In *43-101F1 Technical Report and related consequential amendments*. The Ontario Securities Commission pp. 7043–7086.

Pearce, T. H. (1968). A contribution to the theory of variation diagrams. *Contributions to Mineralogy and Petrology, 19*(2), 142–157.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Piercey, S. J. (2014). Modern analytical facilities 2. A review of quality assurance and quality control (QA/QC) procedured for lithogeochemical data. *Geoscience Canada, 41*, 75–88.

Reed, R., & Marks, R. J., II. (1999). *Neural smithing: supervised learning in feedforward artificial neural networks*. MIT Press.

Rodushkin, I., Engström, E., Pontér, S., & Pennisi, M. (2022). Elemental stable isotope assessment of groundwater contamination: Recent developments. *Current Opinion in Environmental Science & Health, 26*, 100330.

Rossi, M. E., & Deutsch, C. V. (2013). *Mineral resource estimation*. Springer.

Sadeghi, B., Grunsky, E., & Pawlowsky-Glahn, V. (2021). Uncertainty quantification. In B. Daya-Sagar, Q. Cheng, J. McKinley, & F. Agterberg (Eds.), *Encyclopedia of mathematical geosciences. Encyclopedia of earth sciences series* (pp. 334–347). Springer.

SAMREC, 2016. The South African code for the reporting of exploration results, mineral resources and mineral reserves (The SAMREC Code). South African Mineral Codes.

Shanahan, T. M., McKay, N., Overpeck, J. T., Peck, J. A., Scholz, C., Heil, C. W., Jr., & King, J. (2013). Spatial and temporal variability in sedimentological and geochemical properties of sediments from an anoxic crater lake in West Africa: Implications for paleoenvironmental reconstructions. *Palaeogeography, Palaeoclimatology, Palaeoecology, 374*, 96–109.

Sharpe, D. R., Klassen, R. A., Russell, H. A. J., Brennand, T. A., Dexter, K., Logan, C. E., & Stepner, D. A. J. (2019). Regional geochemical survey of earth materials in southern Ontario: A geochemical baseline for environmental and human health protection. *Geological Survey of Canada*. https://doi.org/10.4095/308373.

Teng, G., Jiang, R., Liu, X., Zhou, F., & Chen, Y. (2023). EARN: Toward efficient and robust JPEG compression artifact reduction. *The Visual Computer*. https://doi.org/10.1007/s00371-023-03008-4.

Tolosana-Delgado, R., Mueller, U., & van den Boogaart, K. G. (2019). Geostatistics for compositional data: An overview. *Mathematical geosciences, 51*(4), 485–526.

Tolosana-Delgado, R., & van den Boogaart, K. G. (2014). Towards compositional geochemical potential mapping. *Journal of Geochemical Exploration, 141*, 42–51.

Tolosana-Delgado, R., van den Boogaart, K. G., & Pawlowsky-Glahn, V. (2011). Geostatistics for compositions. In V. Pawlowsky-Glahn & A. Buccianti (Eds.), *Compositional data analysis: Theory and applications* (pp. 73–86). Wiley.

Ulrich, R. N., Guillermic, M., Campbell, J., Hakim, A., Han, R., Singh, S., Stewart, J. D., Román-Palacios, C., Carroll, H. M., De Corte, I., & Gilmore, R. E. (2021). Patterns of element incorporation in calcium carbonate biominerals recapitulate phylogeny for a diverse range of marine calcifiers. *Frontiers in Earth Science, 9*, 641760.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research, 11*(12), 3371–3408.

Xiong, Y., & Zuo, R. (2016). Recognition of geochemical anomalies using a deep autoencoder network. *Computers & Geosciences, 86*, 75–82.

Xiong, Y., & Zuo, R. (2021). Robust feature extraction for geochemical anomaly recognition using a stacked convolutional denoising autoencoder. *Mathematical Geosciences, 54*, 623–644.

Yang, Y., Lv, H., Chen, N., Wu, Y., Zheng, J., & Zheng, Z. (2021). Local minima found in the subparameter space can be effective for ensembles of deep convolutional neural networks. *Pattern Recognition, 109*, 107582.

Zajac, I. S. (2015). John Jambor's contributions to the mineralogy of the Strange Lake peralkaline complex, Quebec-Labrador. *Canada. Canadian Mineralogist, 53*(5), 885–894.

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

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

Zhang, S. E., Nwaila, G. T., Agard, S., Bourdeau, J. E., Carranza, E. J. M., & Ghorbani, Y. (2023a). Big geochemical data through remote sensing for dynamic mineral resource monitoring in tailing storage facilities. *Artificial Intelligence in Geosciences, 4*, 137–149.

Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., & Ashwal, L. D. (2021b). Machine learning-based prediction of trace element concentrations using data from the Karoo large igneous province and its application in prospectivity mapping. *Artificial Intelligence in Geosciences, 2*, 60–75.

Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., Ghorbani, Y., & Carranza, E. J. M. (2023b). Deriving big geochemical data from high-resolution remote sensing data via machine learning: Application to a tailing storage facility in the Witwatersrand goldfields. *Artificial Intelligence in Geosciences, 4*, 9–21.

Zhang, S., Xiao, K., Carranza, E. J. M., Yang, F., & Zhao, Z. (2019). Integration of auto-encoder network with density-based spatial clustering for geochemical anomaly detection for mineral exploration. *Computers & Geosciences, 130*, 43–56.

Zhao, B., Zhang, D., Tang, P., Luo, X., Wan, H., & An, L. (2023). Recognition of multivariate geochemical anomalies using a geologically-constrained variational autoencoder network with spectrum separable module–A case study in Shangluo District China. *Applied Geochemistry, 156*, 105765.

Zhu, W., Mousavi, S. M., & Beroza, G. C. (2020). Seismic signal augmentation to improve generalization of deep neural networks. *Advances in Geophysics, 61*, 151–177.