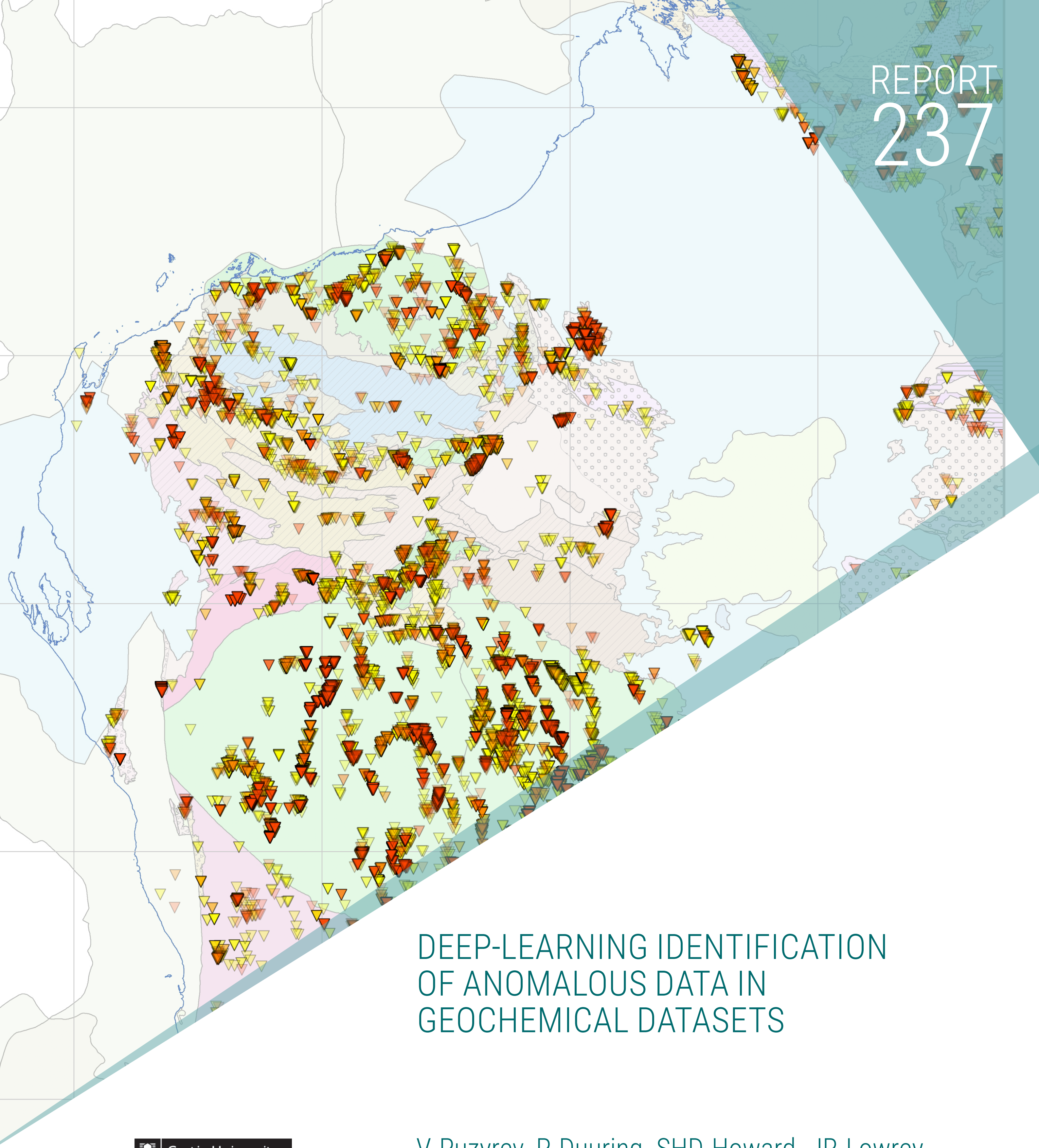
REPORT 237



DEEP-LEARNING IDENTIFICATION OF ANOMALOUS DATA IN GEOCHEMICAL DATASETS

Curtin University
OIL AND GAS INNOVATION CENTRE

V Puzyrev, P Duuring, SHD Howard, JR Lowrey, WR Ormsby, D Purnomo, D Then and J Thom

Government of **Western Australia**
Department of **Mines, Industry Regulation and Safety**

Geological Survey of Western Australia

Government of **Western Australia**
Department of **Mines, Industry Regulation and Safety**

REPORT 237

DEEP-LEARNING IDENTIFICATION OF ANOMALOUS DATA IN GEOCHEMICAL DATASETS

V Puzyrev*, P Duuring, SHD Howard, JR Lowrey, WR Ormsby, D Purnomo,
D Then and J Thom

*School of Earth and Planetary Sciences and Curtin University Oil and Gas Innovation Centre, Curtin University, Perth WA 6102

PERTH 2023

Geological Survey of Western Australia

MINISTER FOR MINES AND PETROLEUM
Hon Bill Johnston MLA

DIRECTOR GENERAL, DEPARTMENT OF MINES, INDUSTRY REGULATION AND SAFETY
Richard Sellers

EXECUTIVE DIRECTOR, GEOLOGICAL SURVEY AND RESOURCE STRATEGY
Michele Spencer

REFERENCE
The recommended reference for this publication is:
Puzyrev, V, Duuring, P, Howard, SHD, Lowrey JR, Ormsby, WR, Purnomo, D, Then, D and Thom, J, 2023, Deep-learning identification of anomalous data in geochemical datasets: Geological Survey of Western Australia, Report 237, 23p.

ISBN 978-1-74168-000-3
ISSN 1834-2280

A catalogue record for this book is available from the National Library of Australia

Grid references in this publication refer to the Geocentric Datum of Australia 1994 (GDA94). Locations mentioned in the text are referenced using latitudes and longitudes. All locations are quoted to at least the nearest 100 m.

About this publication

In this study, a set of deep-learning methods was applied to the harmonized surface and drillhole Western Australian Mineral Exploration (WAMEX) database to identify (and replace) potential spurious occurrences in the data and estimate missing analyte values wherever possible. The methodology is entirely data-driven and, after the corresponding networks have been trained, allows the results to be obtained instantly. Deep-learning methods delivered good results at modest computational cost and, contrary to many other statistical methods, required no manual feature engineering. The results demonstrate the efficacy of this approach for the different types of geochemical data included in the WAMEX database.

Disclaimer

This product uses information from various sources. The Department of Mines, Industry Regulation and Safety (DMIRS) and the State cannot guarantee the accuracy, currency or completeness of the information. Neither the department nor the State of Western Australia nor any employee or agent of the department shall be responsible or liable for any loss, damage or injury arising from the use of or reliance on any information, data or advice (including incomplete, out of date, incorrect, inaccurate or misleading information, data or advice) expressed or implied in, or coming from, this publication or incorporated into it by reference, by any person whosoever.

Published 2023 by the Geological Survey of Western Australia
This Report is published in digital format (PDF) and is available online at <www.dmirs.wa.gov.au/GSWApublications>.

Further details of geoscience products are available from::
First Floor Counter
Department of Mines, Industry Regulation and Safety
100 Plain Street, EAST PERTH WESTERN AUSTRALIA 6004
Telephone: +61 8 9222 3459   Email: publications@dmirs.wa.gov.au
www.dmirs.wa.gov.au/GSWApublications

Cover photograph: A statewide display of all likely spurious WAMEX geochemical data, identified by deep-learning methods

Contents

Appendices

Figures

Tables

Deep-learning identification of anomalous data in geochemical datasets

V Puzyrev*, P Duuring, SHD Howard, JR Lowrey, W Ormsby, D Purnomo, D Then and J Thom

Abstract

The Western Australian Mineral Exploration (WAMEX) database contains geochemical data provided to the Geological Survey of Western Australia (GSWA) in digital format by the exploration and mining industry. The WAMEX database is known to contain a significant amount of spurious data, including errors in unit reporting and incorrect assignment of analytes brought about mainly by errors in post-analysis data reporting and, in some cases, due to low accuracy of the chosen analytical technique. Significant time and cost challenges exist in manually identifying and correcting these issues. In this study, a set of deep-learning methods was applied to the harmonized surface and drillhole WAMEX datasets to identify (and replace) potential spurious occurrences in the data and estimate missing analyte values wherever possible. The method is entirely data-driven and, after the corresponding networks have been trained, allows the results to be obtained instantly. Deep-learning methods delivered good results at modest computational cost and, contrary to many other statistical methods, required no manual feature engineering. The results demonstrate the efficacy of the method for the different types of geochemical data included in the WAMEX database (i.e. surface vs drillhole sample media, and different laboratory analytical methods). The deep neural networks-based estimation approach may be particularly useful when applied to geographical regions in Western Australia where geochemical datasets are incomplete and access to new samples, or reanalysis of existing samples, is inhibited by time, physical access and cost constraints. The predicted values for analytes may benefit mineral explorers by indicating new regions of exploration interest, or simply by highlighting gaps in collected data — the latter informing future data collection strategies that reduce levels of uncertainty and exploration risk.

KEYWORDS: Data processing, drillhole data, geochemistry, neural networks, mineral exploration

Introduction

This Report presents the results of a three-month research project completed by Curtin University and the Geological Survey of Western Australia (GSWA) on deep-learning methods applied to the identification of spurious geochemical data in datasets submitted by companies to GSWA as part of their statutory mineral exploration reporting obligations under the Western Australia *Mining Act 1978*.

The submitted reports and any accompanying datasets are stored in the Western Australian Mineral Exploration (WAMEX) database. Downhole and surface geochemical datasets submitted in digital format (typically from the late 1990s) are imported into a separate database, the Mineral Drillhole Database (MDHDB), with minimal quality control for checking of locations and harmonization of projection datum. The datasets are held in the MDHDB as a series of tables for each type of sample observation, for example, geology, mineralogy, alteration, geochemistry. GSWA releases a public version of the MDHDB from which confidential data have been excluded.

Geochemistry data in the MDHDB include assays and metadata from: a) surface rock chip; b) surface stream sediment; c) surface shallow drillhole; d) surface soil; and e) drillhole samples. The majority of the geochemical analyses are carried out by commercial laboratories and include a wide range of analytical techniques and analyte names. In this project, only publicly released data from the MDHDB was used, therefore the majority of the data used were reported to GSWA more than five years ago. References to WAMEX in the remainder of this report refer to geochemistry data in the MDHDB.

The WAMEX (MDHDB) database contains a significant amount of spurious geochemistry data, including errors in unit reporting and incorrect assignment of analytes brought about mainly by errors in post-analysis data reporting and, in some cases, due to low accuracy of the chosen analytical technique (e.g. portable XRF-derived data commonly have lower accuracy than the majority of the laboratory-based analytical techniques). Some quality control measures were applied during data submission; however, the WAMEX data include at least half a billion analysed samples, varying between single and multi-element analyses. The proportion of spurious data and metadata is estimated to be up to 10%. Time and cost challenges exist in manually identifying and correcting these issues, which involves the identification of the samples in original reports and replacing errors with their true values.

* School of Earth and Planetary Sciences and Curtin University Oil and Gas Innovation Centre, Curtin University, Perth WA 6102

As part of the 2020–21 Accelerated Geoscience program (AGP), GSWA attempted to improve the usefulness of WAMEX geochemical data by making extracts of the database more internally consistent and interrogable. This involved developing a workflow to harmonize the original WAMEX near-surface and drillhole data (the method is described by Duuring et al., 2021), resulting in modified versions of the data with unified analyte fields and consistent unit conventions. However, the harmonized datasets still include spurious data. GSWA published a harmonized version of the WAMEX near-surface geochemical data (Duuring et al., 2021) that is accessible through the DMIRS Data and Software Centre (DASC) <https://dasc.dmirs.wa.gov.au/>. Similarly, a harmonized version of the WAMEX drillhole data may be sourced through GSWA's dedicated web portal <https://wamexgeochem.net.au/>. This study uses public versions of the WAMEX near-surface (surface rock chip, surface stream sediment, surface shallow drillhole, and surface soil datasets) and WAMEX drillhole geochemical data. As both datasets are continually added to, the data used for this project was accessed on 1 April 2021.

Machine learning (ML) methods have been receiving increasing attention in the geoscience community in recent years. The main reason for this is the recent rise of deep-learning (DL) methods that rely on the use of deep neural networks (DNN) that allow unprecedented performance levels in various tasks such as classification, regression, and clustering (Aljalbout et al., 2018; Chalapathy and Chawla, 2019; LeCun et al., 2015). Several attempts have been made to apply ML/DL methods to geochemical data analysis. Most of the approaches use either shallow neural networks or methods such as support vector machines, regression trees and random forests. The latest developments in DL have only been marginally applied to this geochemical data analysis. The main differences between modern DNN and earlier neural networks, besides the obvious increase in hidden layers, are the activation functions commonly employed, the way weights in the hidden neurons are initialized, and the methods of regularization to prevent overfitting. DL algorithms also do not rely on human expertise as much as traditional ML methods. Multiple hidden layers of DNN (usually in the range of tens, sometimes hundreds) allow these models to automatically learn hierarchical feature representations of data with multiple levels of abstraction. This makes DNN a powerful tool to identify anomalies and reveal patterns in large-scale data.

The aim of this research project was to apply the recently developed DL models to identify spurious data and, where possible, estimate missing values within the five harmonized WAMEX geochemical datasets: surface rock chip, surface stream sediment, surface shallow drillhole, surface soil, and drillhole data. The DL techniques can be used to search for hidden dependencies in the geochemical data, and may assist mineral exploration targeting. However, it is important to appreciate that the learning is based on correlations and dependencies between analytes, independent of sample locations and analytical methods. Spurious samples are defined as samples with predicted analyte values that are very different from their measured values. Although tabulated versions of the ML geochemical data have not been publicly released because of their experimental nature, a list of samples was reported to GSWA at the conclusion of the project for further checking and for manual correction to the WAMEX database.

Project data

WAMEX datasets

At the beginning of the project, GSWA provided the following publicly available versions of five harmonized WAMEX geochemical datasets extracted from the MDHDB: a) surface rock chip; b) surface stream sediment; c) surface shallow drillhole; d) surface soil; and e) drillhole data. Datasets a) to d) were obtained from the GSWA harmonized datasets as available on the DMIRS Data and Software Centre. The harmonized drillhole data (dataset e) was obtained from the GSWA's dedicated web portal <https://wamexgeochem.net.au/>.

Company analyte column headings as supplied in submitted datasets were matched to standard analyte names in a match table. Scripts were run on the original WAMEX database to harmonize the company analyte names to the matched standard analyte names, and to recalculate assay values in the various company-supplied units of measure to a standard unit. Most of the samples had been analysed for a range of elements and, to a lesser extent, element oxides, by a variety of analytical techniques at commercial and, in some cases, in-house laboratories.

Data structure

The five harmonized WAMEX geochemical datasets included 54 086 488 samples. Only 7 517 170 were surface samples and the remaining 46 569 318 were drillhole samples (Figure 1 and Table 1). The format of the surface and drillhole databases is slightly different, as described below.

Each sample in the four surface datasets contained 19 information fields, such as sample identity (ID), A-number, coordinates and 124 geochemistry fields, namely:

- 77 analytes that were reported as elements (e.g. Mn) and not as element oxides (e.g. MnO). These analytes are referred to below as main elements. Loss On Ignition (LOI) values were included.
- 42 element oxides that were redundant because they correlated with the main elements (e.g. MnO was disregarded because it correlated with Mn). Exceptions to this rule exist when one of the paired values was missing, or if numerical errors were present.
- 5 columns where no data were reported, that is, there were only null values for these fields in all five databases. These columns were Fe_2O_3, FeO, H_2Oneg, H_2Opos and LABnegNR.

Drillhole datasets contained 13 information fields and 158 geochemistry fields per sample. The increase in the latter is mostly due to a larger number of element oxides reported in drillhole samples (up to 72). However, the number of main elements was the same as in the surface data (n=77).

Table 1 lists the number of samples in each dataset and the corresponding proportions of non-zero values. Although the surface stream sediment was the smallest dataset, it had the highest proportion of non-zero values (i.e. it has the highest proportion of useful information). The drillhole data had the least number of analytes reported.

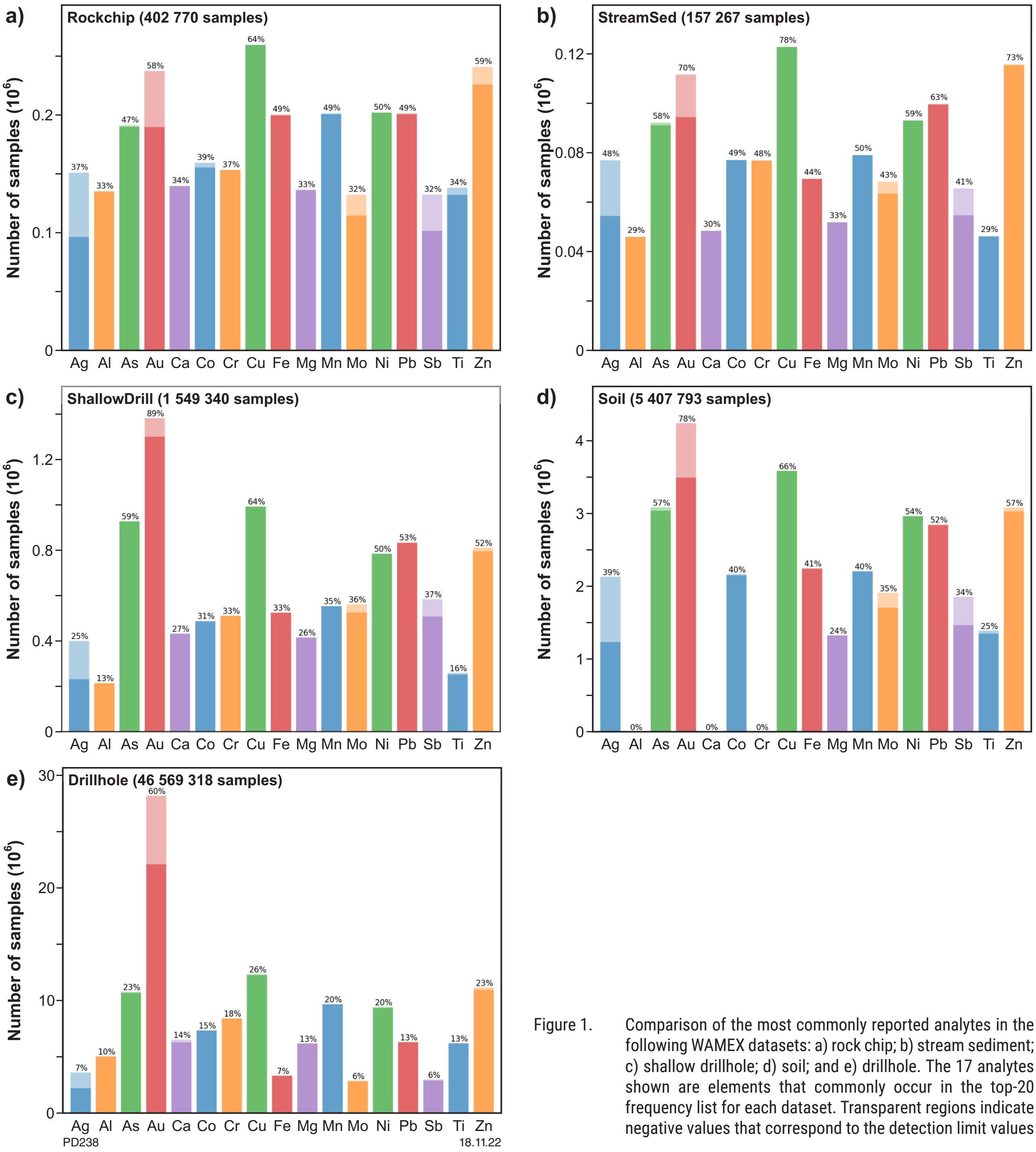


Figure 1. Comparison of the most commonly reported analytes in the following WAMEX datasets: a) rock chip; b) stream sediment; c) shallow drillhole; d) soil; and e) drillhole. The 17 analytes shown are elements that commonly occur in the top-20 frequency list for each dataset. Transparent regions indicate negative values that correspond to the detection limit values

Table 1. Number of total samples and nonzero samples for the five WAMEX datasets (original data, including Fe_2O_3, FeO, H_2Oneg, H_2Opos, and LABnegNR columns).

Dataset	*Number of samples*	*Percentage of nonzero values**
Surface rock chip	402 770	18.26
Surface stream sediment	157 267	25.98
Surface shallow drillhole	1 549 340	15.27
Surface soil	5 407 793	14.99
Drillhole	46 569 318	6.86

NOTE: *After error codes, large positive and negative values have been identified.

Most of the main elements were reported in parts per million (ppm). In the four surface datasets, Platinum Group Elements (PGE) were reported in parts per billion (ppb), whereas LOI, sulfur, and most oxide values were reported in percent.

Table 2 describes each individual subset of the drillhole data. Four subsets, namely water, mining techniques, large diameter, and costean were not considered in this study because they did not satisfy the quality and dataset size requirements for DL methods. The three largest subsets (reverse circulation, aircore and diamond drillhole) and the subset with the highest information content (sonic) were analysed individually. The remaining six subsets of the drillhole samples (percussion, vacuum, rotary mud, auger, rotary air blast, unknown) were merged into an "Others" dataset. Due to the high number of samples with only one analyte reported (e.g. gold) or only under three analytes reported, we used an entry rule for the drillhole samples to be included in the DNN analysis. Namely, we chose samples with no less than five reported analytes (excluding gold, which has poor correlations with other elements).

The nine harmonized WAMEX datasets used in this project are presented below:

- Surface:
  - rock chip
  - stream sediment
  - shallow drillhole
  - soil
- Drillhole:
  - reverse circulation
  - aircore
  - diamond drillhole
  - sonic
  - others (percussion, vacuum, rotary mud, auger, rotary air blast, unknown)

Figure 1 shows the frequency of the commonly reported analytes in each of the five harmonized WAMEX datasets. Overall, the most common analyte reported in WAMEX is gold (despite it only being the third most common analyte in the rock chip and stream sediment datasets). The other most frequently reported elements are the base metals (usually in the order of Cu, Zn, Pb, Ni) and arsenic. Gold, silver, antimony, and, to a lesser extent, cobalt and molybdenum have many values reported as small negative values (corresponding to the detection limit values). We can also observe three individual signatures in the analyte frequency, namely: a) rock chip and stream sediment data were somewhat similar; b) shallow drillhole and drillhole data were very similar; and c) soil data were most different of the five WAMEX datasets (e.g. many metals are not reported). These relationships between datasets are directly linked to the type of sample collected and the laboratory analysis method.

Data pre-processing

In this section, we describe the different steps of data cleaning applied to the datasets before they were delivered to DNN and were standardized.

Error codes

WAMEX datasets have two existing conventional error codes, namely –9999 to indicate null values and –6666 used for samples with values greater than 100%. During the data pre-processing step, it was discovered that some other geochemical values appeared unnatural (e.g. –99990000, –5555, –8888). Although these spurious values represent a very small proportion of the total data for each analyte, they were removed to ensure correct data bounds. All unconventional error codes were treated as obviously spurious data and were replaced by Not a Number (NaN) values similar to the conventional error codes. The full list of error codes is given in Appendix 1.

Table 2. Description of WAMEX drillhole subsets

CODE	*Drillhole type*	*Total number of samples*	*Number of samples included in the analysis*	*Comments*
AC	Aircore	9 257 867	2 167 156	Analysed as an individual dataset
PERC	Percussion	282 371	28 647	Merged into 'Others' dataset
WAT	Water	11 789	–	Excluded from this study
VAC	Vacuum	426 023	157 396	Merged into 'Others' dataset
RM	Rotary mud	1 086 378	5859	Merged into 'Others' dataset
MT	Mining techniques	2790	–	Excluded from this study
AUG	Auger	104 312	29 143	Merged into 'Others' dataset
SON	Sonic	15 140	9133	Analysed as an individual dataset
RAB	Rotary air blasting	4 680 142	1 031 279	Merged into 'Others' dataset
UNKN	Unknown	907 359	83 396	Merged into 'Others' dataset
DD	Diamond drillhole	8 564 058	2 255 821	Analysed as an individual dataset
RC	Reverse circulation	21 209 940	8 605 312	Analysed as an individual dataset
LD	Large diameter	11 332	–	Excluded from this study
COST	Costean	9817	–	Excluded from this study

Element oxide issues

The WAMEX surface and drillhole datasets included up to 42 and 72 element oxide values, respectively. These values were not used in this study if the corresponding main analyte was also reported (e.g. MnO values were not used if their equivalent Mn values were reported). All element oxide – main element pairs were checked for inconsistencies by running correlation tests and checking the ratios of the corresponding elements. For the majority of such pairs, the ratio between the two numbers was constant (as listed in Appendix 2) and coincide with reported element-to-stoichiometric oxide conversion factors (e.g. conversion factors reported on university websites, such as <www.jcu.edu.au/advanced-analytical-centre/resources/element-to-stoichiometric-oxide-conversion-factors>). However, it is noted that some pairs had mismatches (ranging from small, which can be attributed to rounding errors, to several orders of magnitude) between the actual and expected oxide values. There were 24 211 such mismatches in the rock chip data, 1092 in the stream sediment data, 9529 in the shallow drillhole data, 178 458 in the soil data, and 50 031 in the drillhole data. Some of the minor mismatches were due to small rounding issues. For example, in some surface WAMEX samples, the coefficient of 1.2448 is used for zinc oxide, whereas in some other cases its reciprocal value (0.803397) is used for calculations (which results in a coefficient of 1.2447146).

Detection limits

Some samples in the WAMEX data had suspiciously small positive analyte values that were significantly smaller than their corresponding detection limits. Although it was not necessary to remove these values, the DL method was found to achieve better prediction accuracy results when these values were substituted with a standard set of corresponding detection limits for each analyte. Thus, in all datasets we used the lowest detection limits reported by either the WACHEM database, or the ALS and ActLabs commercial laboratories (their respective detection limits are summarized in Appendix 3).

Large positive values

In some very rare cases, reported analyte values were suspiciously large and even reached 1 000 000 ppm. Although such concentrations may not be precise, we decided to keep these values to preserve potentially naturally occurring anomalous element values in the data. PGE values higher than 1 000 000 ppb were removed from the DNN input because it was decided that they were unreasonably large to be naturally occurring.

Negative values

Besides the error codes mentioned above, the WAMEX data included many other negative numbers that mostly corresponded to the detection limit values. These negative numbers (other than conventional and unconventional error codes) were replaced by half of their value, and were reported as positive numbers. However, when these values had a suspiciously large absolute value (e.g. –1000000 or –5000000 ppm, which is obviously spurious), they were replaced by NaN values in the DNN input. The threshold values for each analyte are listed in Appendix 3.

Standardization

The original data *y* was standardized by subtracting the corresponding mean *μ* and dividing by the standard deviation *σ*:

$$z = \frac{y - \mu}{\sigma} \qquad (1)$$

This procedure was performed for each analyte of each dataset independently.

Method

Machine-learning models

We employed DL models that learn mapping from independent variables (input data) to dependent variables (output data). As the geochemical data involves mostly pure numerical input values without spatial or temporal dependence within one sample, we chose a fully connected neural network. The input features were processed by multiple dense layers, which enabled the creation of a hierarchy of feature detection and allowed the model to capture small-scale and large-scale features in the data. An important benefit of this DL approach was that no manual feature extraction or selection was needed as the most representative features were automatically learned during the training process.

The hyperparameters of the network included the variables that determined its structure, such as the number of dense layers, the number of neurons at each layer, activation function employed, dropout values, and the variables that determined the training process (e.g. learning rate and number of epochs). Hyperparameters are typically chosen to avoid underfitting and overfitting on the data. For example, too deep neural networks can achieve very good performance in training although overfit the training data and thus generalize poorly to new inputs (Goodfellow et al., 2016). This can be partially mitigated by using regularization techniques such as dropout (Srivastava et al., 2014). A practical measure of accuracy is how well the algorithm performs on data that it has not seen before. This ability is called generalization and it often determines the real-world performance of a method.

In the following examples, hyperparameter tuning was performed using the k-fold cross-validation (Kohavi, 1995) and grid search process. In the k-fold cross-validation technique, the training set is split into k different groups. The algorithm iterates through these groups and, at each iteration, uses one of them as the validation set and the remaining k-1 groups as the training set to train the model. The process is repeated k times. In this study, we employed 95% of each WAMEX dataset as the data given to the DL models for training using the k-fold cross-validation (k = 5).

The remaining 5% of the data (chosen randomly and called the test set below) was not shown to the DL models and was used to estimate the accuracy of the trained networks on new data (assuming that the solution is unknown for the test data). All the performance metrics reported below, unless otherwise stated, were calculated on these test datasets that were separate from the data used for training. Spurious data identification and missing values estimation were performed for both training and test subsets of each harmonized WAMEX dataset.

We used networks that had 10 to 12 hidden layers and 128 to 256 neurons in each layer. A leaky rectified linear unit (LReLU) was used as the activation function at each hidden layer. The rectified linear unit (ReLU) activation function and its variants (Xu et al., 2015) are a common choice for convolutional networks and allow high accuracy in many other types of networks. The main benefits of ReLUs are sparsity, reduced likelihood of vanishing gradient, better convergence performance, and cheap computational cost. The LReLU variant was chosen instead of standard rectified linear units to avoid the so-called ReLU problem (Maas et al., 2013), which arises because standard ReLU functions return zero values from negative inputs – potentially resulting in a large part of the network that stops learning. A dropout of 0.05 – 0.15 is applied after each layer, except the last one. These choices were made based on an extensive series of experiments for each scenario.

For optimization, we employed the Adam algorithm (Kingma and Ba, 2014), which is a common choice due to its computational efficiency and fast convergence. The majority of the standard optimization algorithms provide similar performance (Puzyrev, 2019). The implementation used the open-source software library TensorFlow (Abadi et al., 2016) and scikit-learn (Pedregosa et al., 2011). The training was performed on two systems equipped with NVIDIA Tesla V100 and P100 GPUs and took several days for the WAMEX datasets. Once trained, the networks can be used on a mid-range laptop/desktop graphics card. For example, the missing data estimation and spurious samples detection procedures reported here were performed on a mid-range GTX 940MX card (released in 2016) and took between several minutes (stream sediment data) to approximately two hours (soil data) to process.

Accuracy metrics

Training a neural network is an optimization problem equivalent to finding the minima of the loss function, which measures the accuracy of the predicted model. Choosing a loss function that leads to the desired behaviour of the method is an important task. Two common metrics that measure accuracy for continuous variables without considering their direction are the mean absolute error (MAE) and mean squared error (MSE). The latter is usually preferred when large errors are undesirable. MSE is a quadratic scoring rule (equivalent to the difference in L_2 norm) that squares the errors before they are averaged, thus giving a relatively high weight to large errors:

$$\mathrm{MSE}(z,\hat{z}) = \frac{1}{N}\sum_{i=1}^{N}\left(z_i - \hat{z}_i\right)^2 \tag{2}$$

Here $\boldsymbol{z}_i$ and $\boldsymbol{\hat{z}}_i$ denote, respectively, the normalized estimated and true values for each sample calculated using Equation (1). The neural network is trained using the root mean squared error (RMSE) as the loss function, which is a common choice for continuous variables.

As additional quality metrics, we used several absolute and relative error metrics. The first one is the mean squared error for denormalized values (i.e. returned back to original range). For convenience, we used the square root. The RMSE metric is calculated as:

$$\mathrm{RMSE}(y,\hat{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2} \tag{3}$$

Here $\boldsymbol{y}_i$ and $\boldsymbol{\hat{y}}_i$ denote, respectively, the denormalized estimated values and true values (i.e. original values of the analyte) for each sample. The RMSE is an absolute error measure that can range from 0 to infinity and is indifferent to the direction of errors. It is also dependent on the magnitude of the variables, which varies significantly for different analytes in geochemical data. Thus, the RMSE is not sufficient for a full evaluation of the accuracy of the results.

A more useful error metric is the relative metric called the symmetric mean absolute percentage error (SMAPE). This metric is independent of the scale and measures the accuracy based on percentage errors as follows:

$$\mathrm{SMAPE}(y,\hat{y}) = \frac{200\%}{N}\sum_{i=1}^{N}\frac{\left|y_i - \hat{y}_i\right|}{\left|y_i\right| + \left|\hat{y}_i\right|} \tag{4}$$

The right-hand denominator here represents the average of the actual value and the forecast, which avoids division by zero when true values are zero. For other advantages of SMAPE over the non-symmetric mean absolute percentage error (MAPE), we refer the reader to Tofallis (2015). For small errors, both MAPE and SMAPE metrics are very similar; see Ghommem et al. (2021) for an example. The SMAPE metric formulated as Equation (4), which is a commonly used formulation, has a lower bound of 0% and an upper bound of 200%. Over-forecasts and under-forecasts are not treated equally (in a linear sense) by the SMAPE metric; however, their ratios are. For example, an overestimation of factor 3 and an underestimation of factor 3 both result in the same SMAPE value (100%).

To quantify the correlation between the predicted and true values, we reported the commonly used coefficient of determination ($\boldsymbol{R^2}$) and (less commonly used although potentially more powerful) concordance correlation coefficient (CCC) proposed by Lin (1989). $\boldsymbol{R^2}$ determines the proportion of the variance in the dependent variables that is predictable from the independent variables and is formulated as:

$$R^2(y,\hat{y}) = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5}$$

where $SS_{res} = \sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)$ is the sum of squares of residuals

and $SS_{tot} = \sum_{i=1}^{N}\left(y_i - \mu\right)^2$ is called the total sum of squares

(modified by the mean of the observed data). $\boldsymbol{R^2}$ is equal to 1 when the predicted values exactly match the actual values and will be zero for a model that always predicts the mean value. $\boldsymbol{R^2}$ can accept negative values.

The CCC measures the concordance between two sets of values to quantify the agreement between them (Lin, 1989). It is formulated as:

$$\mathrm{CCC}(y,\hat{y}) = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2+\sigma_{\hat{y}}^2+\left(\mu_y-\mu_{\hat{y}}\right)^2} \qquad (6)$$

where ρ is Pearson's correlation coefficient between the two variables. CCC cannot exceed the absolute value of ρ between the same datasets and also ranges from –1 to 1, with perfect agreement at 1. CCC often produces more meaningful scores compared to R^2 and other correlation measures. There is no strict agreement on how to interpret CCC values, although a common approach is to consider positive values less than 0.2 as having no concordance and values greater than 0.8 as having good concordance.

Results

This section reports the main statistics for spurious data (anomaly) detection and missing data predictions, and provides several examples.

In these experiments, we trained the estimating and anomaly detecting DNN separately for each analyte in each of the nine harmonized WAMEX datasets. Combining data from different datasets, for example merging the rock chip and soil data, led to a significant decrease in prediction accuracy. Analytes that had less than 100 non-zero values for a given set were excluded from the procedure due to lack of data, namely: C, CO_2, F, I and Ir.

The soil dataset lacked data for Al, Ba, Be, Bi, Br, Ca, Cd, Ce, Cl, Cr, Cs, Dy and Er.

In the digital outputs containing spurious data, and in the figures shown below, we used the following three-colour system to indicate predictability for each analyte. Green values refer to high predictability because they have SMAPE errors below 50% (equal to an overestimation of factor 5/3 or an underestimation of factor 3/5). Yellow values have moderate predictability with SMAPE errors between 50% and 100% (factor 3 for overestimation or underestimation). SMAPE values above 100% are red to indicate low predictability. True values used in SMAPE calculations, and shown on the plots, are the values reported in the harmonized WAMEX datasets and may include errors. For example, the most common reasons for large errors arose in the harmonization process where units were sometimes incorrectly assigned (e.g. ppb reported as ppm) and due to incorrect mapping of analytes.

Prediction accuracy – examples from stream sediment data

Table 3 shows the accuracy metrics, Equations (2) to (6), for each analyte of the test subset of the stream sediment data. Although it was the smallest dataset (157 267 samples), the stream sediment dataset had the highest information content and achieved the best accuracy metrics for most analytes. In terms of analyte accuracy, the observed distribution for the stream sediment dataset is qualitatively similar to the other WAMEX datasets. The rare earth elements (REE) achieved the highest accuracy on the test data due to very high correlations within this group (i.e. an unknown value of one REE can be estimated from several known values of other REE). Some analytes (Br, Cl, I and, to a lesser extent, C and LOI) were rarely reported in the stream sediment data, which results in larger SMAPE errors. Gold was often reported; however, its predictability is normally the lowest among other common analytes. Silver, As and Hg were the analytes with the next lowest predictability.

Figure 2 shows a typical graphical example of a DNN-based estimation, in this case showing predicted nickel values vs the harmonized nickel values (labelled True) from the test stream sediment dataset, together with their relative errors (i.e. the corresponding SMAPE plot). For nickel, the predictability is quite high across all ranges of values, which vary by almost seven orders of magnitude. There is no observable bias in the nickel estimation. The SMAPE plot shows a common pattern of having low errors in the middle zone, where the majority of the samples are located, and higher errors at both the low- and high-end groups of the samples.

In Figures 3 and 4, we show similar estimations for stream sediment samples; however, we consider two elements that have a lower predictability than base metals, namely silver and arsenic. Despite larger error metrics, the DL method could reliably distinguish low-content samples from high-content ones (which also means that severe spurious data anomalies are almost absent in this dataset; for example, only a few samples were mismatched by two orders of magnitude or more, while for other datasets, this number was higher). The SMAPE plots demonstrate that there were larger differences in the ranges where there were fewer training data. Vertical clusters seen in silver (Fig. 3) and, to a lesser degree, in arsenic (Fig. 4) distributions, correspond to the actual data being rounded (e.g. 0.01, 0.005, 0.001). Both these elements have a similar number of spurious data anomalies and almost the same average SMAPE (32%). Correlation metrics were significantly better for arsenic, which can be explained by the higher continuity of its actual values (i.e. fewer reported values were rounded), and the overall data distribution.

An example of an analyte with high predictability is given in Figure 5 for neodymium, which achieved the lowest error metrics for the stream sediment data. Such high prediction accuracy is explained by very strong correlations in the REE group. Most of these elements achieved similar high accuracy. The largest errors among REE were typically for scandium, yttrium, lanthanum, cerium and lutetium (although some other REE had strong spurious data anomalies, particularly in the drillhole data).

Comparison of the prediction accuracy between the harmonized WAMEX datasets

In Table 4, we report the average prediction accuracy for all analytes from the four surface harmonized WAMEX datasets and five categories of harmonized WAMEX drillhole datasets.

Table 3. Accuracy metrics, based on Equations (2) to (6), for the stream sediment data measured on the 5% randomly extracted test data

ID	Analyte	Normalized MSE	RMSE	CCC	R^2	SMAPE (%)
1	Ag	2.45E-03	8025.374	0.198	0.158	32.194
2	Al	1.62E-04	567766.958	0.966	0.932	7.909
3	As	1.14E-03	2055.971	0.879	0.793	31.918
4	Au	4.29E-03	262.900	0.270	0.131	58.412
5	B	4.50E-03	850.288	0.784	0.655	29.789
6	Ba	4.32E-04	10533.831	0.888	0.791	18.623
7	Be	2.39E-04	120.821	0.400	0.260	11.191
8	Bi	6.88E-04	676.912	0.457	0.317	24.451
9	Br	6.51E-04	14078.728	0.917	0.871	28.065
11	Ca	9.47E-04	339074.874	0.884	0.750	22.042
12	Cd	7.33E-04	106.390	0.658	0.593	17.152
13	Ce	4.80E-05	169055.702	0.985	0.974	8.248
14	Cl	8.71E-03	5535.807	0.246	0.127	59.954*
15	Co	3.10E-04	18217.137	0.873	0.798	18.438
17	Cr	3.24E-04	1841121.717	0.980	0.964	21.722
18	Cs	2.00E-04	61.241	0.967	0.937	14.105
19	Cu	5.65E-04	5598.242	0.658	0.514	24.753
20	Dy	3.00E-05	262.591	0.993	0.987	4.880
21	Er	2.00E-05	53.603	0.999	0.997	4.567
22	Eu	6.70E-05	31.498	0.994	0.988	7.417
24	Fe	6.91E-04	212247.644	0.936	0.879	15.329
25	Ga	7.00E-05	118.976	0.971	0.940	7.699
26	Gd	3.30E-05	16684.453	0.983	0.969	5.748
27	Ge	1.48E-04	17.449	0.998	0.995	6.993
28	Hf	7.70E-05	227.063	0.975	0.940	8.971
29	Hg	7.91E-04	10148.933	0.852	0.805	34.770
30	Ho	3.50E-05	113.250	0.948	0.921	4.606
31	I*	1.51E-03	22.086	0.853	0.426	35.527
32	In	1.11E-03	5.216	0.794	0.739	18.157
33	Ir	5.65E-04	4065.947	0.653	-0.052	7.174
34	K	3.63E-04	48316.390	0.916	0.809	12.778
35	La	4.50E-05	34982.742	0.993	0.984	8.276
36	Li	2.07E-04	431.109	0.945	0.900	14.795
37	LOI	2.04E-03	18.494	0.668	0.502	23.350
38	Lu	3.76E-04	10059.679	0.886	0.751	12.545
39	Mg	2.95E-04	721668.024	0.958	0.917	19.050
40	Mn	4.01E-04	774760.265	0.876	0.802	20.946
41	Mo	6.69E-04	13083.665	0.990	0.978	26.215
42	Na	6.82E-04	307497.939	0.835	0.563	20.097
43	Nb	3.11E-04	2144.255	0.949	0.910	17.478
44	Nd	1.60E-05	3049.869	0.990	0.981	3.962
45	Ni	3.19E-04	5231.087	0.880	0.799	19.861
46	Os	2.19E-04	0.058	0	-inf**	0.735
47	P	3.42E-04	70006.885	0.997	0.994	16.606
48	Pb	4.79E-04	2044.994	0.721	0.606	22.167
49	Pd	2.60E-03	11894.992	0.945	0.896	29.880
50	Pr	1.50E-05	720.498	0.994	0.986	4.110
51	Pt	1.36E-03	143578.322	0.014	0.012	22.759
52	Rb	2.11E-04	628.979	0.990	0.980	11.165
53	Re	2.50E-04	17.303	0.958	0.932	9.388
54	Rh	9.50E-05	0.269	0	-inf**	2.620
55	Ru	2.71E-03	14207.117	0.512	0.272	27.200
56	S	1.92E-03	53.215	0.989	0.975	27.823
57	Sb	3.56E-04	10103.006	0.972	0.954	24.348
58	Sc	1.28E-04	94.521	0.982	0.965	9.284
59	Se	6.79E-04	18.055	0.967	0.933	8.382
60	Si	7.60E-04	784680.080	0.929	0.864	15.145
61	Sm	2.90E-05	524.003	0.993	0.986	4.724
62	Sn	2.84E-04	207.300	0.721	0.488	15.108
63	Sr	3.04E-04	1213.624	0.964	0.930	15.326
64	Ta	2.79E-04	195.815	0.950	0.909	17.309
65	Tb	2.00E-05	71.689	0.990	0.978	4.490
66	Te	3.23E-04	3570.517	0.548	0.506	11.269
67	Th	1.84E-04	14639.574	0.970	0.941	15.496
68	Ti	3.93E-04	747602.823	0.845	0.631	19.912
69	Tl	2.03E-04	15452.215	0.844	0.793	13.878
70	Tm	2.60E-05	361.019	0.987	0.972	5.601
71	U	1.30E-04	6799.181	0.949	0.909	14.818
72	V	1.87E-04	7422.136	0.825	0.740	12.803
73	W	8.56E-04	23084.175	0.840	0.691	23.931
74	Y	8.80E-05	905.807	0.997	0.993	8.013
75	Yb	4.30E-05	77.562	0.995	0.989	6.246
76	Zn	4.12E-04	205178.194	0.979	0.958	22.444
77	Zr	1.04E-04	16158.367	0.898	0.853	12.199

Red indicates the analytes with the largest errors and poorest predictability. Light-blue denotes the rare earth elements (which typically have high predictability). Light-green shows other elements with a moderate predictability. Elements #10 (C), #16 (CO_2) and #23 (F) were excluded from the estimation due to the very low number of samples where these elements are reported

* Cl and I are in the list but are not highlighted with red due to insufficient samples (especially compared to red Ag, As, Au and Hg)

** Os and Rh have their CCC equal to zero and R^2 minus infinity due to all actual values in the test dataset being the same

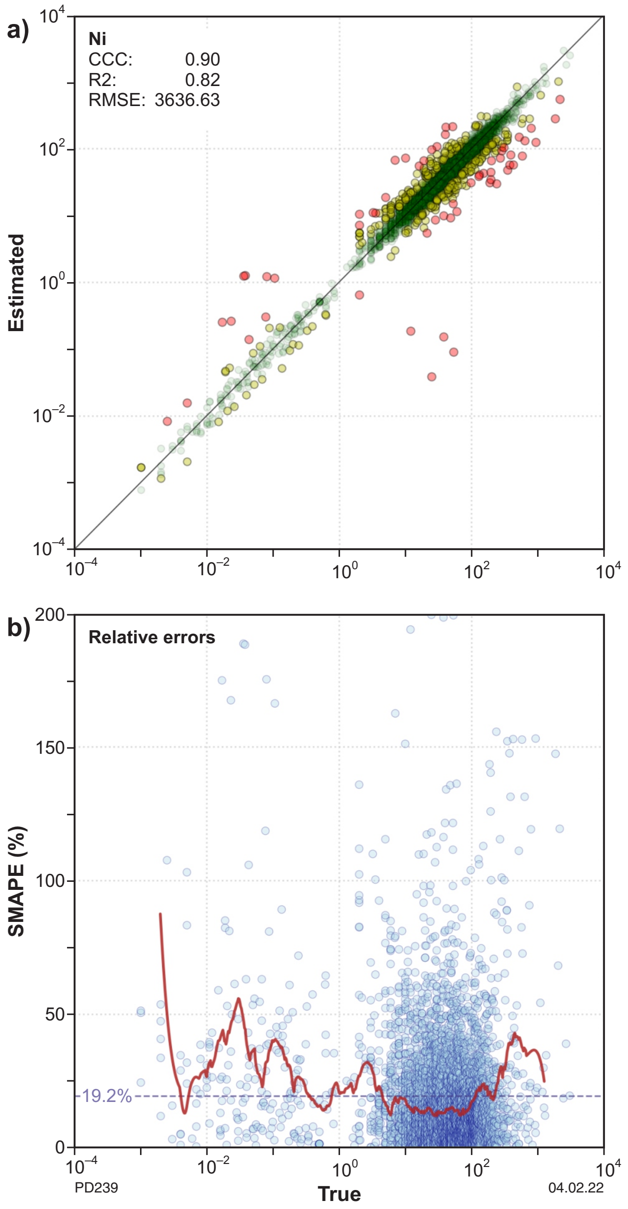


Figure 2. Nickel predictions: a) estimated Ni values vs the corresponding true (harmonized) values for each sample in the test stream sediment dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively; b) SMAPE for individual samples (blue circles) and smoothed-over averaged sample values (red lines). The horizontal dashed line shows the corresponding average SMAPE over the entire test dataset. The smoothed red curve is obtained by 1D interpolation of individual samples to 1000 points evenly spaced on a log scale and a subsequent smoothing with a Savitzky-Golay filter (Savitzky and Golay, 1964) with a window length of 5 and a degree 2 polynomial (Virtanen et al., 2020)

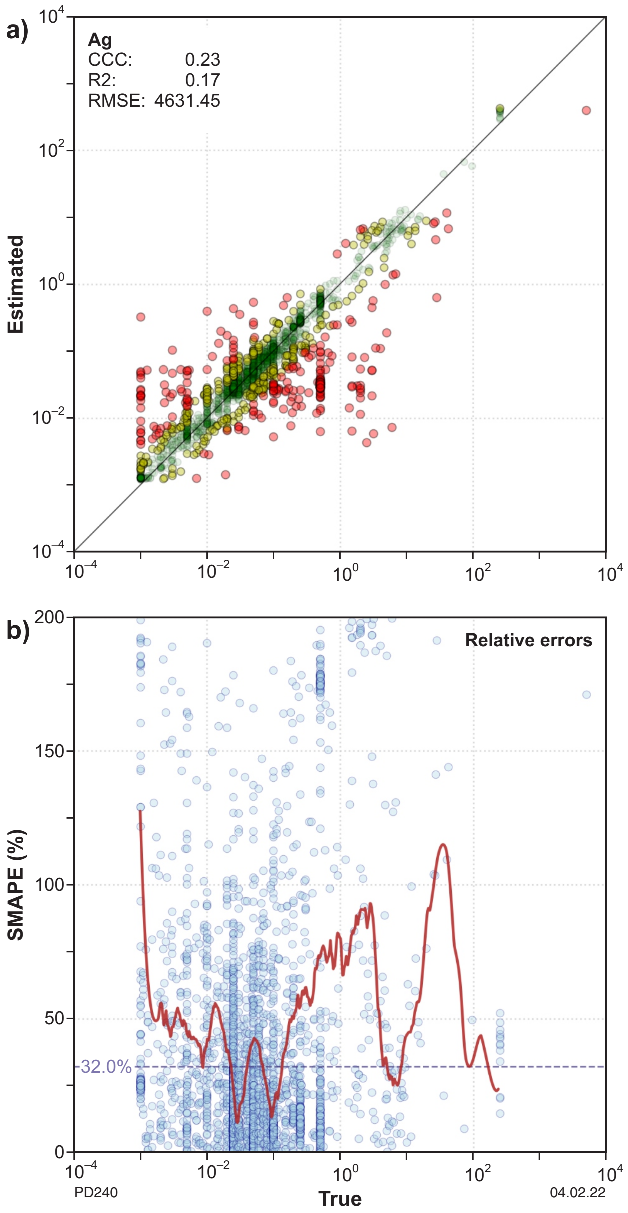


Figure 3. Silver predictions: a) estimated Ag values vs their corresponding true (harmonized) values for each sample in the test stream sediment dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively; b) SMAPE for individual samples (blue circles) and smoothed-over averaged sample values (red lines). The horizontal dashed line shows the corresponding average SMAPE over the entire test dataset

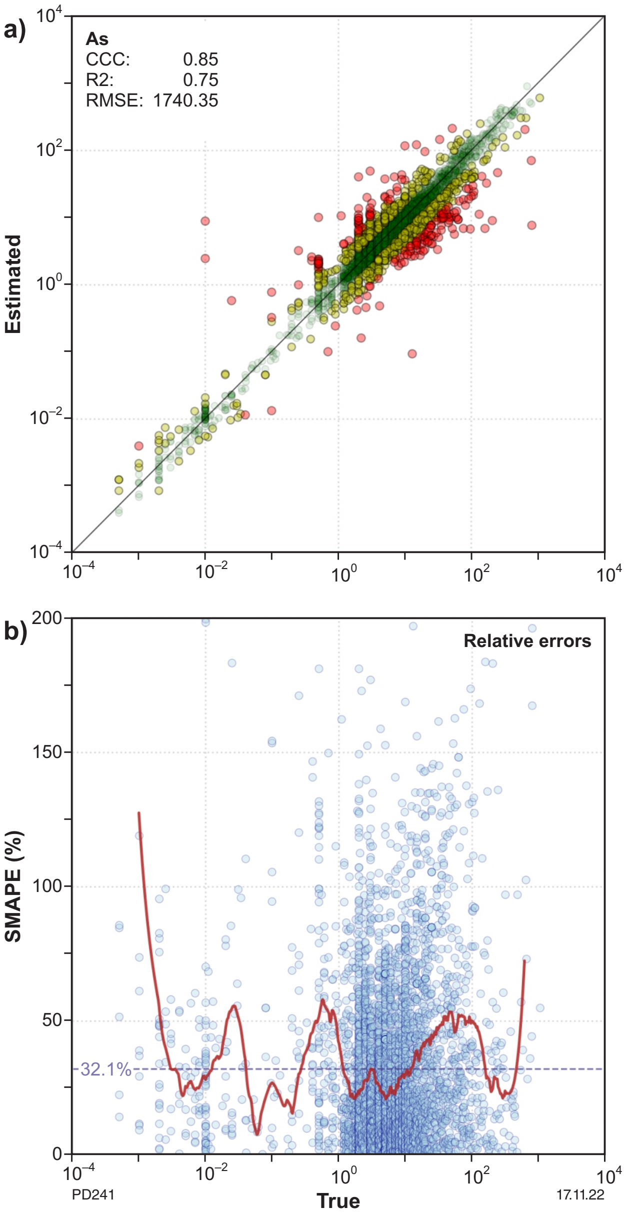


Figure 4. Arsenic predictions: a) estimated As values vs their corresponding true (harmonized) values for each sample in the test stream sediment dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively; b) SMAPE for individual samples (blue circles) and smoothed-over averaged sample values (red lines). The horizontal dashed line shows the corresponding average SMAPE over the entire test dataset

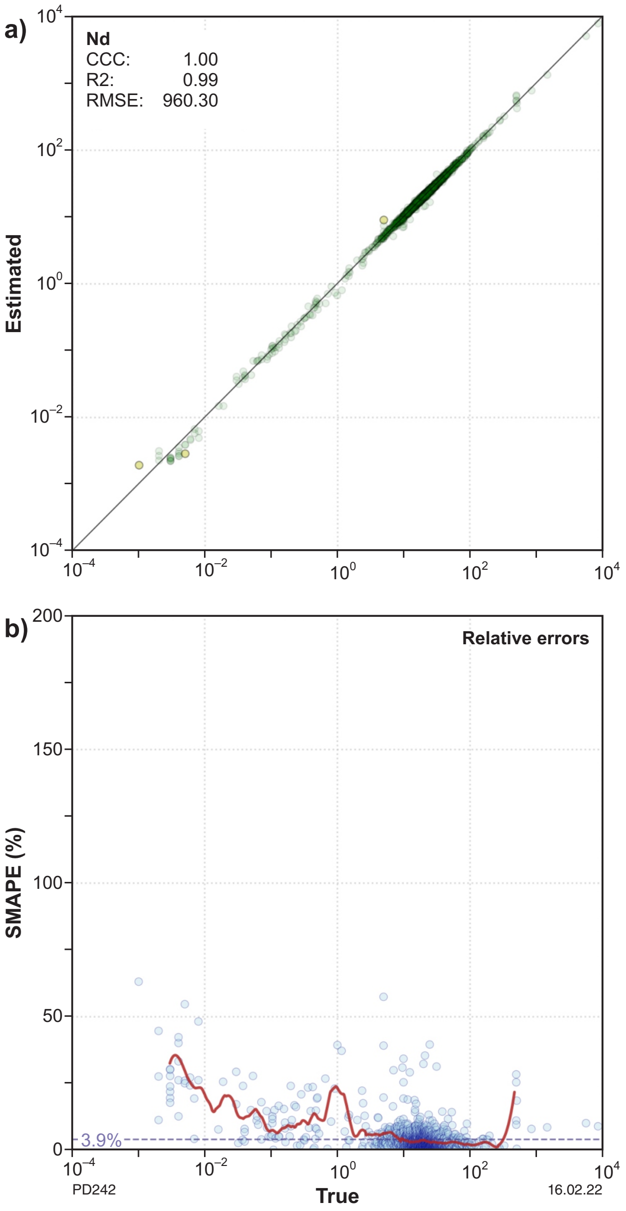


Figure 5. Neodymium predictions: a) estimated Nd values vs their corresponding true (harmonized) values for each sample in the test stream sediment dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively; b) SMAPE for individual samples (blue circles) and smoothed-over averaged sample values (red lines). The horizontal dashed line shows the corresponding average SMAPE over the entire test dataset

Table 4. Average CCC, R^2 and SMAPE metrics calculated on the test subsets of the four surface and five categories of drillhole WAMEX datasets

Test dataset	*Average CCC*	*Average R^2*	*Average SMAPE*
Surface rock chip	0.728	0.514	**31.315**
Surface stream sediment	0.827	0.754	**16.884**
Surface shallow drillhole	0.849	0.708	**17.798**
Surface soil	0.764	0.639	**20.790**
Drillhole sonic	0.823	0.737	**26.891**
Drillhole diamond	0.811	0.712	**25.829**
Drillhole aircore	0.755	0.674	**31.075**
Drillhole reverse circulation	0.804	0.719	**30.447**
Drillhole 'Others'	0.764	0.674	**31.031**

NOTE: Non-weighted values are used; hence, each analyte has an equal contribution to the average score. Analytes that do not have enough data for prediction are not included in the average score

The surface stream sediment and surface shallow drillhole datasets had the highest prediction accuracy, indicated by all three error metrics used, whereas the surface soil dataset had slightly higher average SMAPE, and lower CCC and R^2 scores. Most spurious data were detected in the surface rock chip, drillhole aircore, and drillhole "Others" datasets. For both drillhole datasets, this can be explained by a lower ratio of reported:missing values in the data (see Table 1). In contrast, the surface rock chip dataset had a higher reporting ratio. Another reason for the poor performance can be that these datasets are not as homogeneous, that is, they display mixing of data obtained from multiple analytical methods (more results from portable XRF analyses, which tend to have poorer accuracy than laboratory-derived data). The highly accurate, low detection-level analytical techniques (typically used for surface soil and stream sediment sampling) combined with the large proportion of fine sample fraction may also contribute to the comparatively low average SMAPE values.

Table 5 shows the prediction accuracy for different groups of elements for the four surface and five drillhole WAMEX datasets. The base metals predictions delivered high accuracy (except for rock chip data and the five drillhole sets where average errors were higher). Precious metals, on average, had higher errors. The possible reasons for this are many negative values (below detection limit) in the reported gold and silver, and poor natural correlation between gold and other elements (i.e. even Ag and Pt are poorly correlated with Au). REE were estimated with very high accuracy (SMAPE; 4 – 6.5%) in the majority of the cases; the only exceptions can be explained by either the small size of the dataset (drillhole sonic; 16.6%) or the heterogeneity of the samples (surface rock chip and drillhole "Others", 10.61% and 9.52%, respectively).

Table 5. Average SMAPE values for three different groups of elements in the test subsets of the four surface and five drillhole WAMEX datasets

Test dataset	*Average SMAPE*		
	Ag/Au/Pt	*Cu/Ni/Pb/Zn*	*REE (6)*
Surface rock chip	53.33	48.75	10.61
Surface stream sediment	37.79	22.31	4.12
Surface shallow drillhole	40.65	25.17	4.04
Surface soil	43.24	28.11	6.52
Drillhole sonic	50.42	29.34	16.60
Drillhole diamond	39.51	42.04	5.69
Drillhole aircore	45.70	41.06	6.32
Drillhole reverse circulation	45.27	45.67	6.36
Drillhole 'Others'	50.27	44.38	9.52

NOTE:Non-weighted values are used within each group

Spurious data detection – examples from WAMEX drillhole data

All potentially spurious samples detected in the harmonized WAMEX datasets (including their training and test parts) were reported in nine reports to GSWA (i.e. one report per dataset). In this section, we provide examples of spurious data. Surface stream sediment data displayed the least number of suspicious elements (which did not form obvious clusters) compared to the other datasets.

Figure 6 shows an example of DNN-based estimations for aluminium in the test drillhole aircore dataset. The accuracy of the predictions was significantly lower for this metal compared to most of the other datasets. For example, the average SMAPE was more than three times higher compared to the surface stream sediment data. Although this can be explained by several factors (e.g. low information content in drillhole data, see Table 1, or more measurement/reporting errors), another interesting feature is worth mentioning. The dataset had a cluster of severely overestimated samples near the detection limit. Such suspicious vertical lines have been observed for other metals, such as barium, lanthanum, manganese and scandium, across the drillhole datasets. Most of these potentially spurious samples originated from relatively few industry reports submitted to the WAMEX database. The common vertical trend demonstrates that the values reported in the database are around the detection limits, whereas the DNN predictions suggest much higher concentrations (see the highlighted vertical line in Fig. 6). The most likely reason for such high mismatch is errors in unit reporting (confirmed for at least a representative subset of example data by comparing the potentially spurious data against their reported values in the original WAMEX reports).

Figures 7 to 9 show similar clusters of potentially spurious data found in the DNN-based estimations for lanthanum, scandium and cobalt in the test drillhole reverse circulation data.

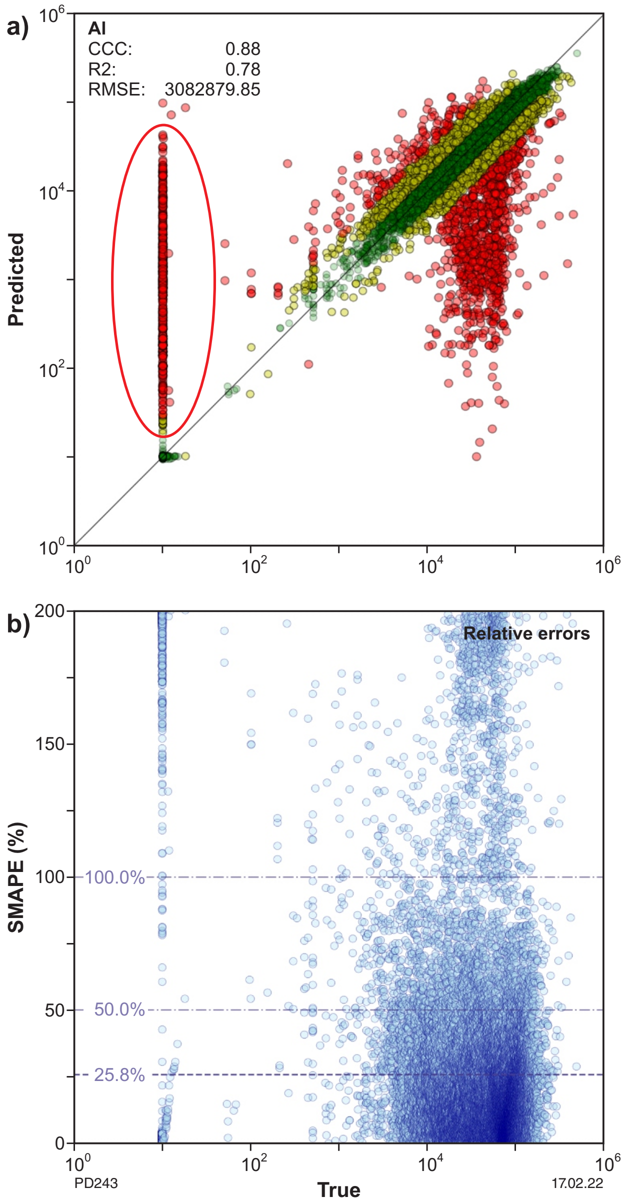


Figure 6. Aluminium predictions: a) estimated Al values vs their corresponding true harmonized values for each sample of the test drillhole aircore dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively. The spurious data region is denoted with the red oval; b) SMAPE for individual samples. The horizontal dashed lines show the corresponding average SMAPE over the entire test dataset

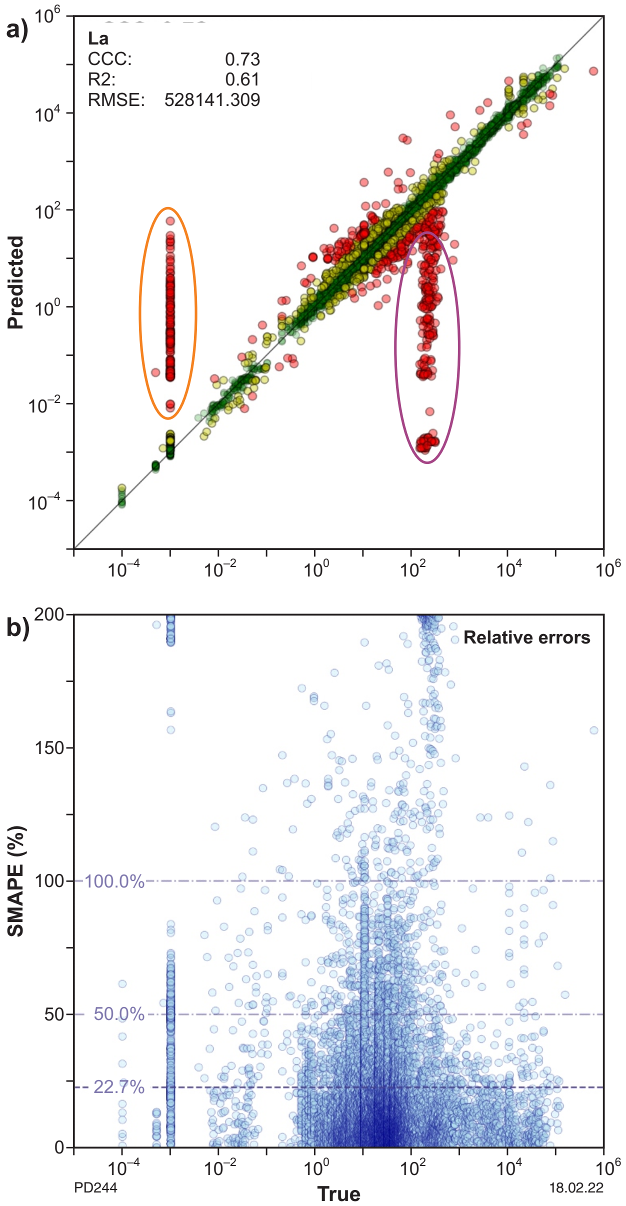


Figure 7. Lanthanum predictions: a) estimated La values vs their corresponding true harmonized values for each sample of the test drillhole reverse circulation dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively. Two spurious data regions are denoted with the orange and purple ovals; b) SMAPE for individual samples. The horizontal dashed lines show the corresponding average SMAPE over the entire test dataset

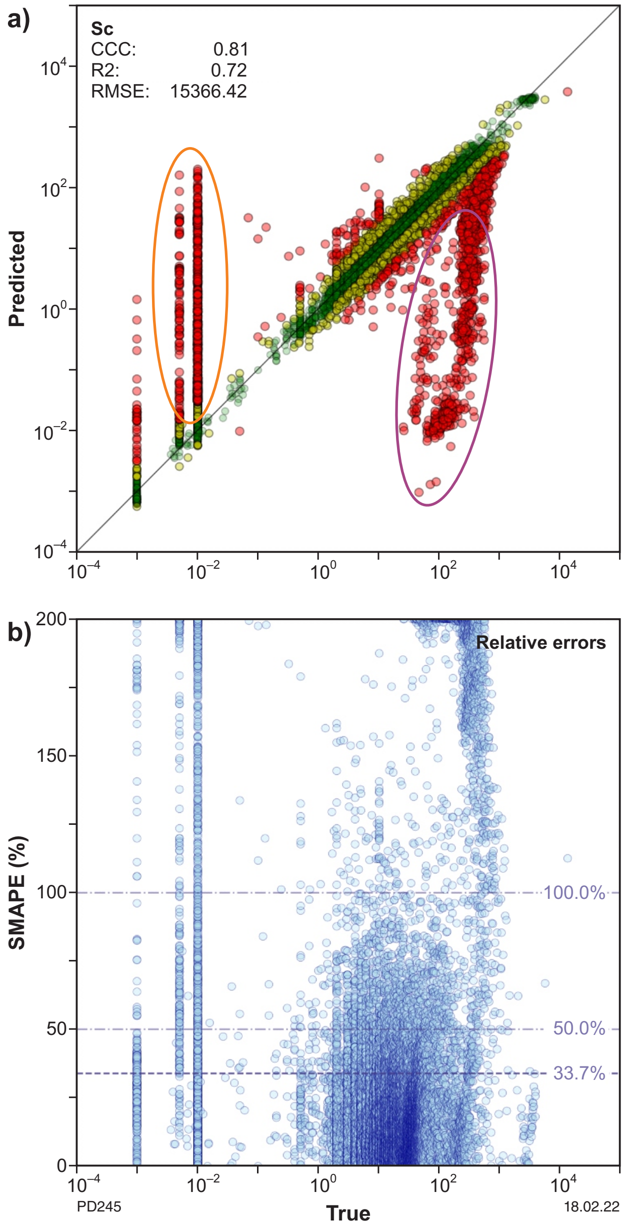


Figure 8. Scandium predictions: a) estimated Sc values vs their corresponding true harmonized values for each sample of the test drillhole reverse circulation dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively. Two spurious data regions are denoted with the orange and purple ovals; b) SMAPE for individual samples. The horizontal dashed lines show the corresponding average SMAPE over the entire test dataset

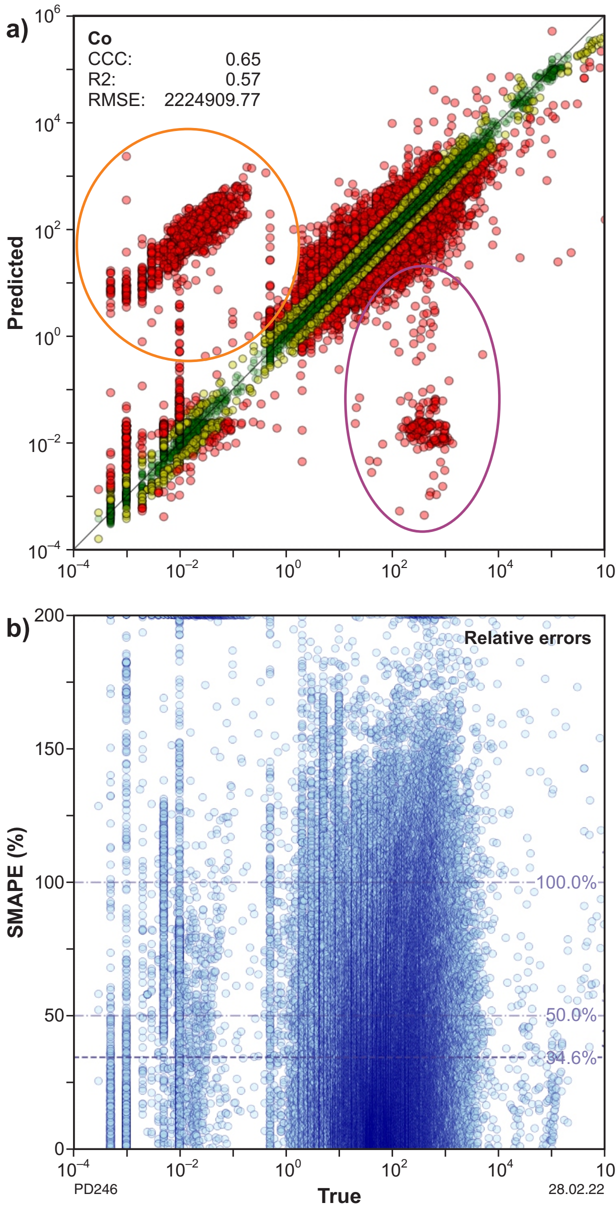


Figure 9. Cobalt predictions: a) estimated Co values vs their corresponding true harmonized values for each sample of the test drillhole reverse circulation dataset. Green, yellow and red denote the 0–50%, 50–100% and 100–200% SMAPE bands, respectively. Two spurious data regions are denoted with the orange and purple ovals; b) SMAPE for individual samples. The horizontal dashed lines show the corresponding average SMAPE over the entire test dataset

These analytes differ from the other analytes in the reverse circulation data due to large clusters of both underestimations and overestimations. In particular, cobalt (Fig. 9) has a collection of samples with the same exploration (A-number) report, with values approximately four orders of magnitude smaller than the estimated values. Such large differences are typically caused by incorrect unit reporting (i.e. percent vs ppm).

Spatial maps

Based on an assessment of all harmonized WAMEX surface and drillhole datasets, Figure 10 shows the spatial distribution of samples in Western Australia that have potentially spurious reported data (i.e. they display large errors in at least one reported analyte). The spurious values for gold (46% of all spurious data) and chlorine (very few samples) are not shown in Figure 10 due to lower estimation confidence for these analytes.

Figure 11 shows the location of potentially spurious samples in cobalt in the test drillhole reverse circulation data (i.e. by plotting the spatial distribution of the two clusters in Fig. 9).

Figures 12 to 15 show several representative examples of spatial maps (for Ni, Li and Ag for both the training and test subsets) that compare the locations of surface and drillhole samples with high predicted values vs known occurrences of high-content samples. Such maps are valuable to explorers because they provide a gap analysis for these targeted elements, possibly indicating areas of higher exploration potential. For each example shown, we chose a minimum analytical threshold denoted as T. Samples with analyte values below T are not included. Analyte values above T are shown in yellow, orange or red depending on their concentration (the exact limits documented in each figure caption).

Figures 12 and 13 show the locations of harmonized WAMEX surface soil samples with estimated high nickel and lithium values, respectively, and compare them to samples with high known values of these analytes. Both these metals show good accuracy in their test data. Nickel is shown using a T of 500 ppm (Fig. 12), whereas for lithium T was 50 ppm (Fig. 13).

In Figure 14, we compare the locations of harmonized WAMEX diamond drillhole samples with higher estimated and reported nickel values. A threshold of 3000 ppm was used for nickel.

In Figure 15, we compare the locations of diamond drillhole samples with estimated high silver concentration with the samples with reported high silver concentration. The T for Ag was 5 ppm.

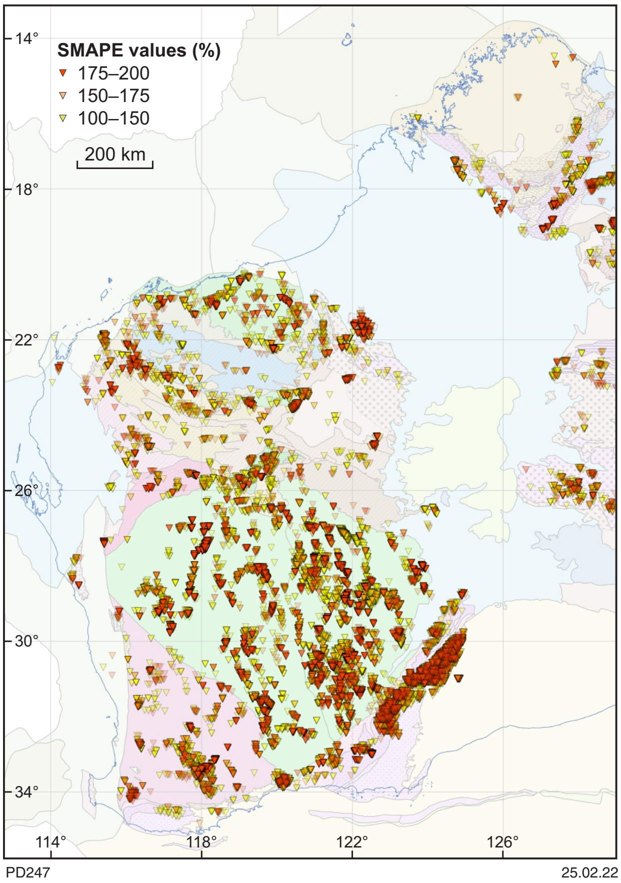


Figure 10. All potentially spurious WAMEX samples (excluding Au and Cl) are shown on GeoVIEW's tectonic units map. Yellow symbols denote samples with SMAPE between 100% and 150%. Orange symbols indicate samples in the 150–175% range, while red symbols denote samples with SMAPE of 175–200%

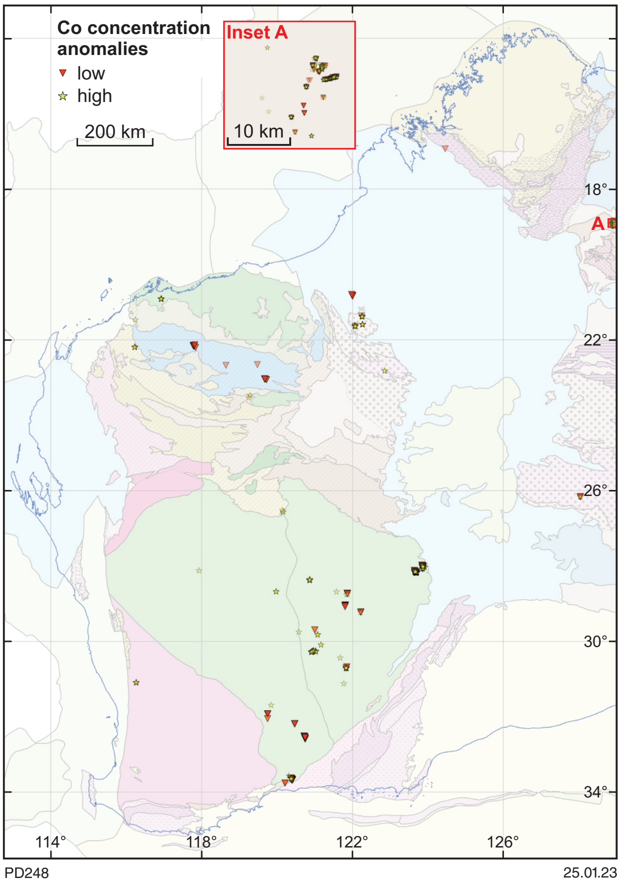


Figure 11. Drillhole reverse circulation WAMEX samples with suspicious cobalt values shown on GeoVIEW's tectonic units map. Triangles and stars denote, respectively, the values that seem to be too low or too high compared to the estimated cobalt content. An inset map is provided of a small area along the Western Australia–Northern Territory border to highlight the issue of overlapping data when viewed at a large scale

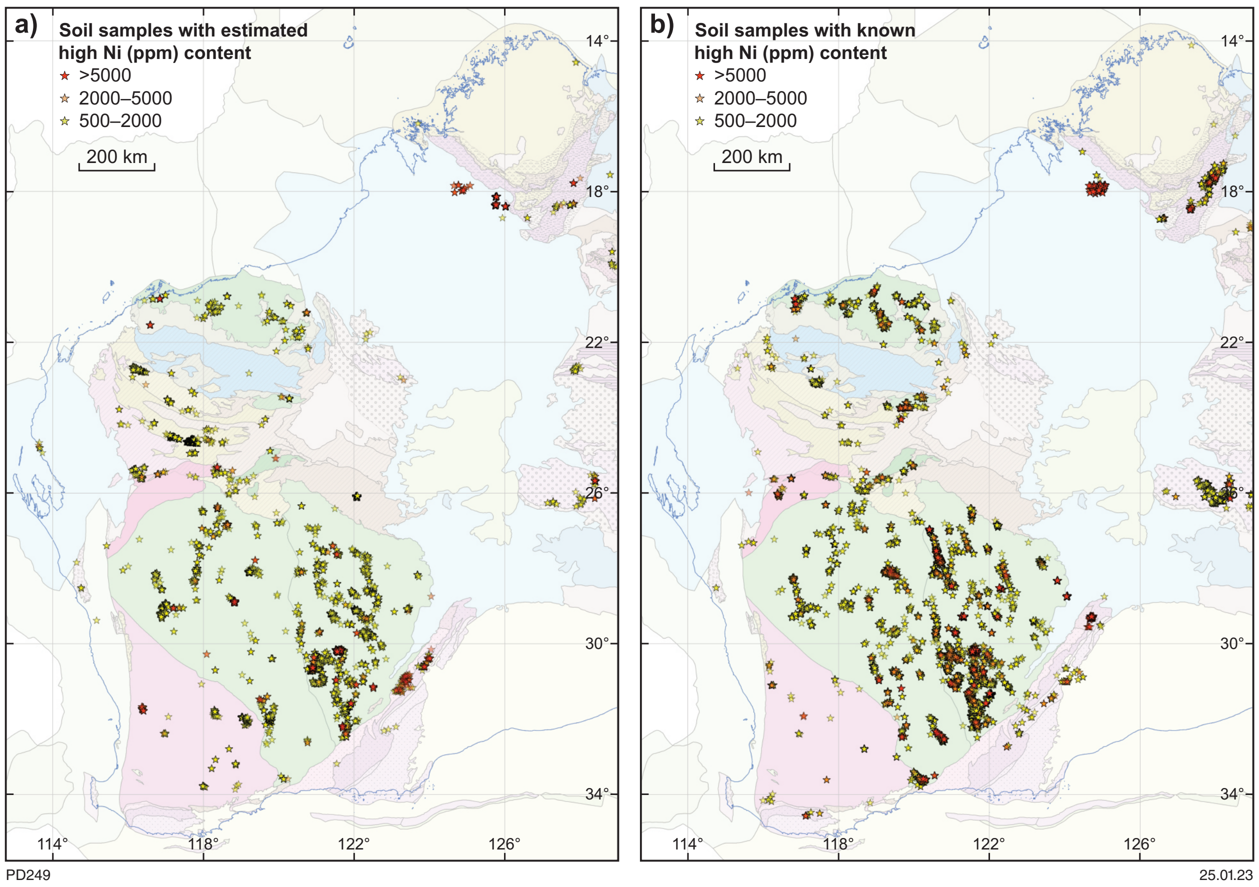


Figure 12. A gap analysis method for identifying prospective areas for nickel exploration: a) WAMEX surface soil samples with estimated high Ni content; b) samples with known high Ni content. Both training and test subsets are shown. Yellow, orange and red symbols denote samples with Ni concentration of 500–2000 ppm, 2000–5000 ppm and above 5000 ppm, respectively

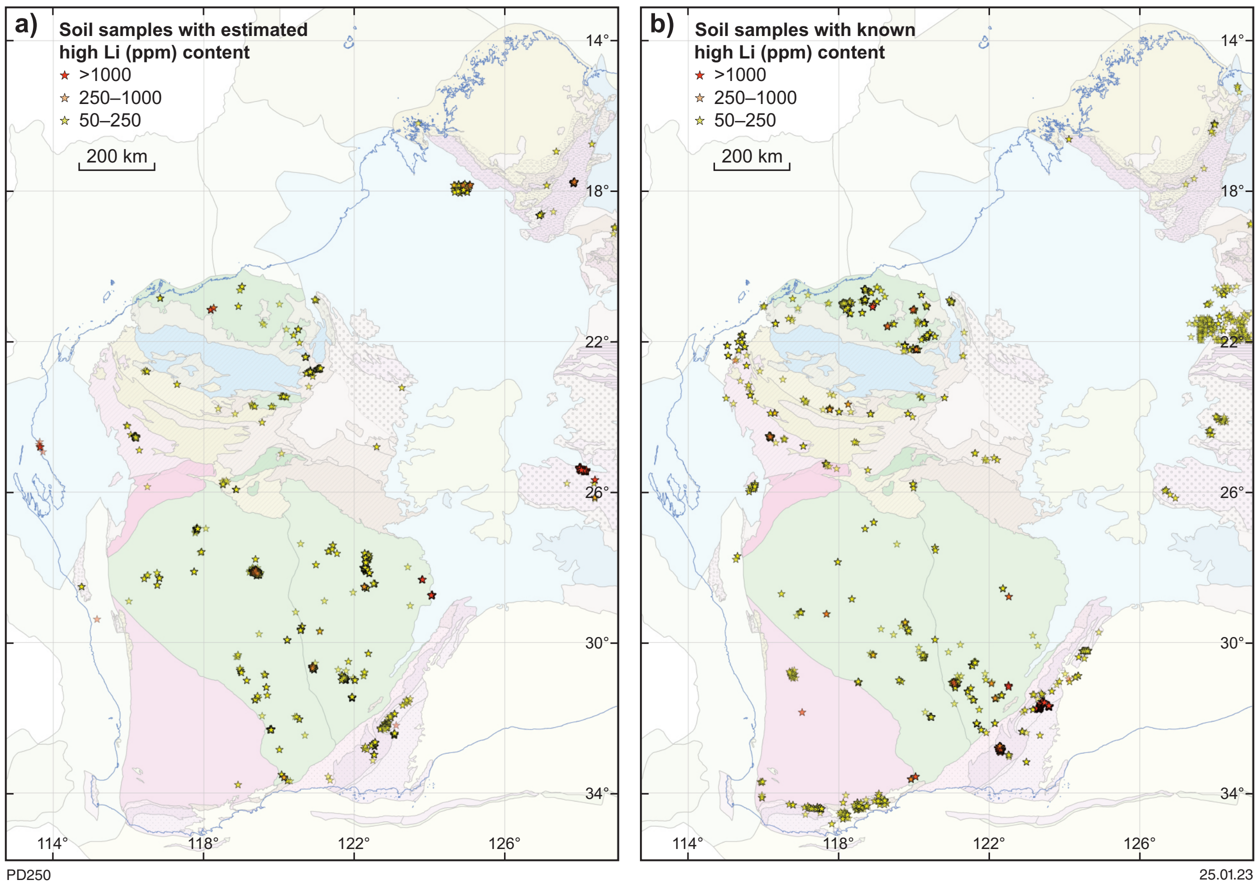


Figure 13. A gap analysis for lithium: a) WAMEX surface soil samples with estimated high Li content; b) surface samples with known high Li content. Both training and test subsets are shown. Yellow, orange and red symbols denote samples with Li concentration of 50–250 ppm, 250–1000 ppm and above 1000 ppm, respectively

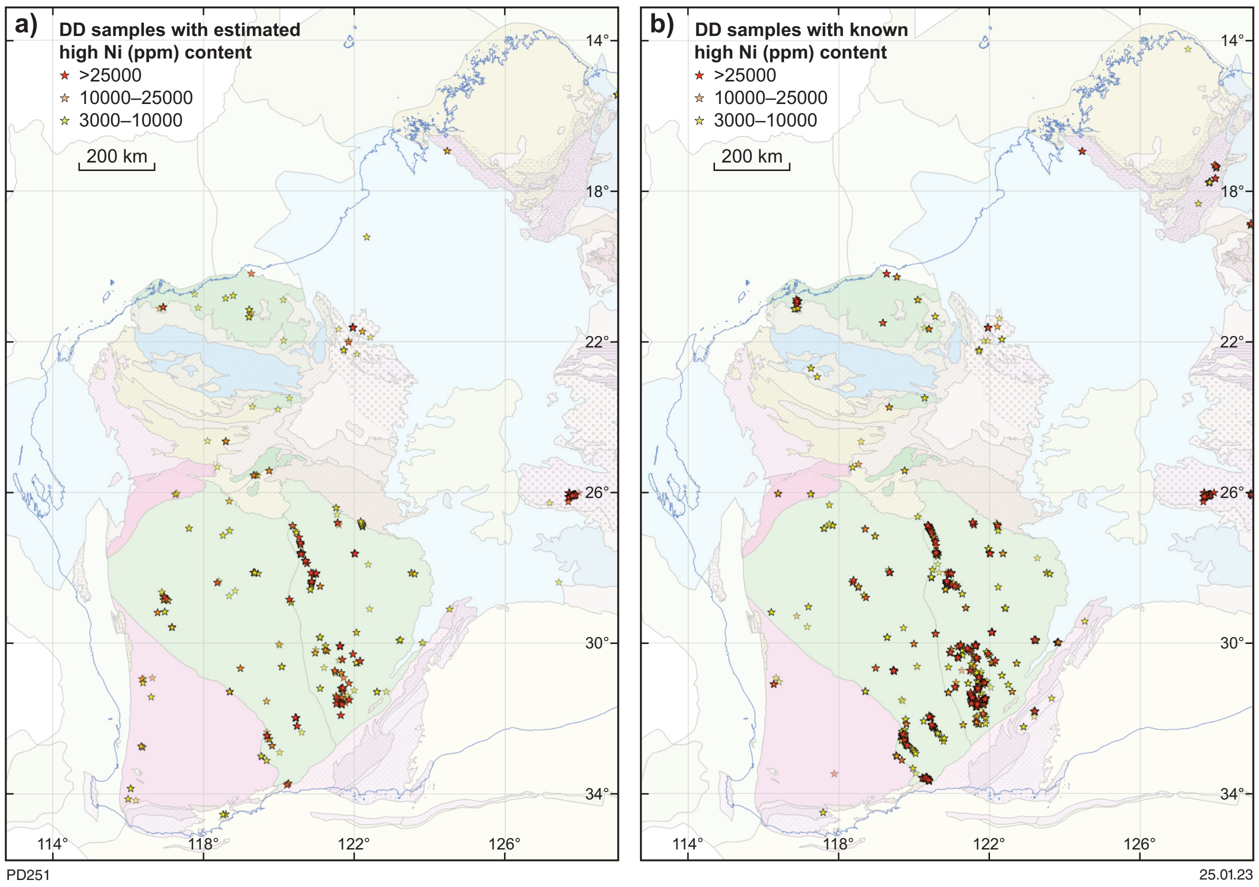


Figure 14. Nickel prospectivity interpreted from diamond drillhole (DD) data: a) WAMEX diamond drillhole samples with estimated high Ni content; b) samples with known high Ni values. Both training and test subsets are shown. Yellow, orange and red symbols denote samples with Ni concentration of 3000–10 000 ppm, 10 000–25 000 ppm and above 25 000 ppm, respectively

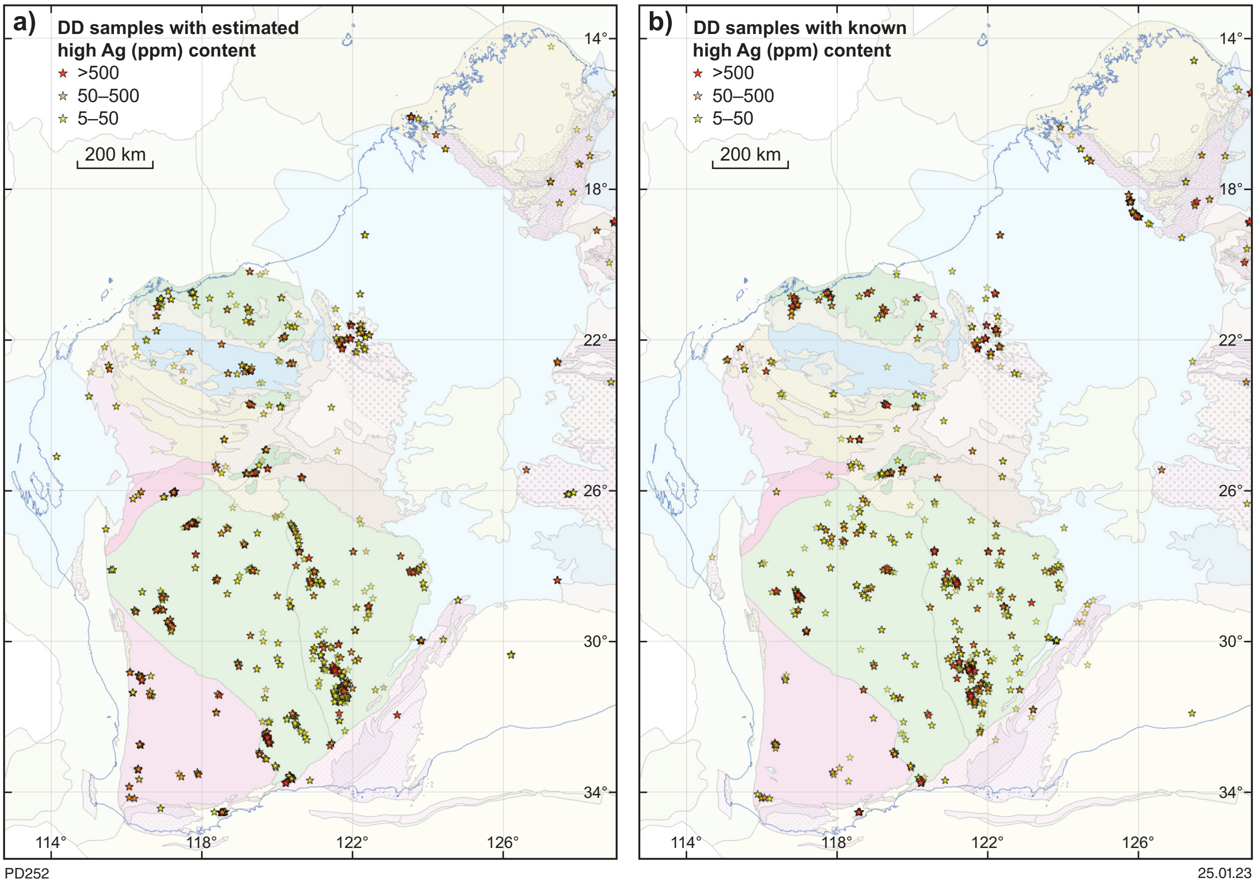


Figure 15. Prospective areas for silver: a) WAMEX diamond drillhole samples with estimated high Ag content; b) samples with known high Ag values. Both training and test subsets are shown. Yellow, orange and red symbols denote samples with Ag concentration of 5–50 ppm, 50–500 ppm and above 500 ppm, respectively

Discussion

We applied a set of DL methods to the harmonized surface and drillhole WAMEX datasets to identify (and replace) potential spurious data and estimate missing analyte values where possible. The method was entirely data-driven and, after the corresponding networks were trained, allowed the results to be obtained instantly. In estimating missing values, we considered each of the analytes in each of the WAMEX datasets separately. The exact configuration of the networks was determined by the optimal validation set error for each scenario. Predicting missing data in a dataset only using data from that dataset was found to be more accurate compared to using data from several datasets.

Figure 16 shows two examples of word clouds for the analytes that were found to have the lowest or highest predictability from known elements across all datasets. Gold (in large red text in Fig. 16a) is poorly correlated with other elements and is thus difficult to estimate reliably. Silver, arsenic and mercury (in orange) have significant prediction errors; however, their order of magnitude is usually predicted correctly. Chlorine is classified as having a low predictability mainly due to the low number of samples. An interesting observation is the absence of platinum, which, despite having a low predictability, was not one of the 10 analytes with the lowest predictability in the five individual datasets (drillhole data was considered as a single dataset in this comparison). A possible reason is the moderately strong correlations of platinum with palladium. The rest of the elements that have low predictability (in green-blue colours) can be estimated without multiple significant errors if the training data is of high quality.

The list of the most reliably predicted elements shown in Figure 16b almost exclusively consists of REE, which is explained by strong correlations within this group. To decrease the influence of these REE, Figure 17 shows a word cloud of the list of elements with the highest predictability, excluding the 17 REE. Under these new constraints, Ga, Hf and Si have the highest predictability. Although Rh, Ir and Os feature in this list, they may have lower predictability compared to the other listed elements in Figure 17 because Rh, Ir and Os are rarely analysed in the WAMEX datasets.

The main limitation for estimating missing data is the quality of the training datasets. An investigation of representative spurious data demonstrated that common causes were harmonization errors and poor accuracy of the input (true) data; for example, pXRF data with poor accuracy were used in training together with data obtained by more accurate laboratory methods. For large databases such as WAMEX, cleaning of the data should be performed as a reiterative process in which obvious spurious data are identified, removed from training, and then the networks retrained.

The results of experiments in this project show that DL methods deliver good results at modest computational cost and, contrary to many other statistical methods, require no manual feature engineering. The results also demonstrate the efficacy of the method for the different types of geochemical data included in the WAMEX database (i.e. surface vs drillhole sample media, and differences between laboratory analytical methods). The DNN-based estimation approach may be particularly useful when applied to geographical regions in Western Australia where geochemical datasets are incomplete and access to new samples, or reanalysis of existing samples, is inhibited by time, physical access and cost constraints. The predicted values for analytes in this DL version of WAMEX data may benefit mineral explorers by indicating new regions of exploration interest, or simply by highlighting gaps in collected data — the latter informing future data collection strategies that reduce levels of uncertainty and exploration risk.

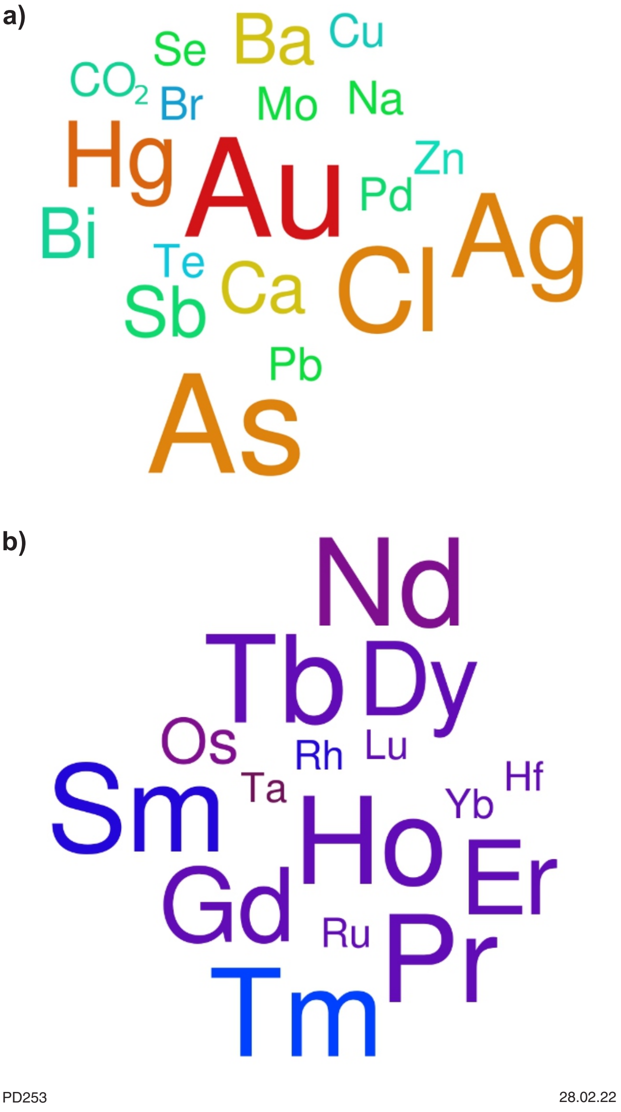


Figure 16. Analytes in the harmonized WAMEX data: a) most difficult to predict; and b) easy to predict. The top 10 elements for each dataset are presented. The size and colour of the analyte represents the frequency of its occurrence in the top 10 (for example, gold and neodymium scored 5 out of 5). In a), red indicates low predictability and blue represents higher predictability

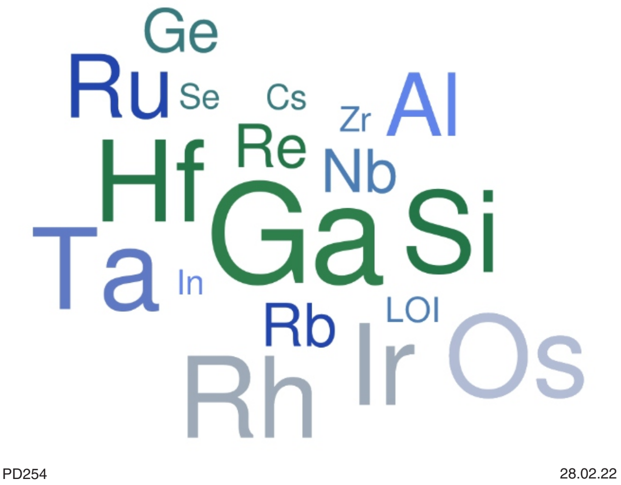


Figure 17. Most reliably predicted analytes in the harmonized WAMEX data excluding the REE. Top 10 elements for each dataset are presented. The size and colour of the analyte represents the frequency of its occurrence in the top 10. Ir, Os and Rh have low support compared to other elements and are intentionally grey

Future outlook

The current project is the first step towards a DL-based system for geochemical data quality assurance and quality control (QA/QC). Such tools are needed because manual spurious data detection requires too much time and human resources. This three-month project explored the feasibility of the proposed approach in spurious data detection and missing values estimation in large-scale WAMEX geochemical data, and found it to be a viable method. A tangible outcome of this project was a list of samples with probable spurious geochemical values. Subsequent validation of this sample list by GSWA has led to the identification of reporting errors, such as incorrect allocation of unit values, in the reported data. Although this manual correction to the WAMEX geochemical data has been applied in some cases, we acknowledge that complete and thorough data validation may be time-consuming. Based on the obtained results, we briefly outline possible directions for future work.

1. **Quality control of database inputs.** The fast prediction capabilities of DNN allow for rapid quality checking of new geochemical data from commercial laboratories. This quality check could be applied as new data are generated from analysed samples, or could be performed retrospectively and applied to all existing data stored by a company. The cleaned WAMEX geochemical data could be used as the premier training set with different sample types selected so that they match the end user's sample types. This quality check could also be applied to all industry-generated geochemical data that are routinely reported and integrated into the WAMEX database. This automated system would allow DMIRS to more easily issue rapid feedback to companies and request corrected data.

2. **Assessment of confidence in missing data estimation and uncertainty quantification.** The quality of missing data prediction is highly dependent on the quality and information content of each individual sample. DNN is a black-box model with decisions sometimes hard to interpret. Therefore, development of a tool that allows quality assessment of our estimations of the missing values for each individual sample will be of great practical use. This is also true for estimating the confidence in the DNN identification of spurious data. Improvements in quality from retraining on an improved harmonized dataset will allow for a better understanding on the causes of quality issues (e.g. training data, inherent process issues, or physical/chemical constraints). Data estimation quality can be further assessed for a specific geological province/mineralization type.

3. **More specialized DL-based tools for spurious data detection.** Currently, we use the fully connected DNN that was developed primarily for the missing data estimation. There are significant potential advantages in exploring novel DL methods developed for spurious data detection, which may lead to improved accuracy and, similar to the previous point, an increase in confidence in the results. Some research has been done in this direction in other fields (Chalapathy and Chawla, 2019); however, not in the geochemical data context.

4. **Rock type classification from geochemistry.** Rock type classification tasks can be efficiently handled with a similar DNN-based approach. Based on our previous experience with the WACHEM database (Puzyrev et al., 2023), such DL estimation of the rock type of a sample from its geochemistry shows good prediction accuracy for the entire database. In some cases, the classification accuracy exceeds 90% for 10 rock classes (e.g. the Sir Samuel–Leonora–Menzies study area with its abundant regolith samples that have a more precise prediction compared to other possible rock type categories; Puzyrev et al., 2023). This can be useful for validating existing datasets or predicting rock types directly from geochemical data.

5. **Spatial assessment of missing data estimations.** Comparison of mineralization trends derived from the existing harmonized WAMEX datasets (Ormsby et al., 2021) with those from the DL missing data prediction could provide insights into the applicability of this approach for targeted exploration.

Conclusions

The WAMEX database is known to contain a significant amount of spurious data, including errors in unit reporting and incorrect assignment of analytes. In this study, a set of DL methods was applied to the harmonized surface and drillhole WAMEX datasets to identify (and replace) potential spurious data and estimate missing analyte values where possible. The method is entirely data-driven and, after the training phase is complete, allows the results to be obtained instantly. The results demonstrate the efficacy of the method in detecting potential spurious entries and estimating missing analyte values for the different types of geochemical samples included in the WAMEX database.

Acknowledgements

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

References

Abadi, M, Barham, P, Chen, J, Chen, Z, Davis, A, Dean, J, Devin, M, Ghemawat, S, Irving, G, Isard, M, Kudlur, M, Levenberg, J, Monga, R, Moore, S, Murray, DG, Steiner, B, Tucker, P, Vasudevan, V, Warden, P, Wicke, M, Yu, Y and Zheng, X 2016, TensorFlow: a system for large-scale machine learning: arXiv.org, p. 265–283, doi:10.48550/arxiv.1605.08695.

Aljalbout, E, Golkov, V, Siddiqui, Y, Strobel, M and Cremers, D 2018, Clustering with deep learning: taxonomy and new methods: arXiv.org, doi:10.48550/arxiv.1801.07648.

Chalapathy, R and Chawla, S 2019, Deep learning for anomaly detection: a survey: arXiv.org, doi:10.48550/arxiv.1901.03407.

Duuring, P, Then, D, Howard, D and Morin-Ka, S 2021, Western Australian near-surface geochemistry, *in* Accelerated Geoscience Program extended abstracts *compiled by* Geological Survey of Western Australia: Geological Survey of Western Australia Record 2021/4, p. 185–186.

Ghommem, M, Puzyrev, V and Najar, F 2021, Deep learning for simultaneous measurements of pressure and temperature using arch resonators: Applied Mathematical Modelling, v. 93, p. 728–744.

Goodfellow, I, Bengio, Y and Courville, A 2016, Deep Learning: MIT Press, Cambridge, USA, 800p.

Kingma, DP and Ba, J 2017, Adam: a method for stochastic optimization: arXiv.org, doi:10.48550/arxiv.1412.6980.

Kohavi, R 1995, A study of cross-validation and bootstrap for accuracy estimation and model selection, *in* Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada, 20–25 August 1995, p. 1137–1145, <www.ijcai.org/Proceedings/95-2/Papers/016.pdf>.

LeCun, Y, Bengio, Y and Hinton, G 2015, Deep learning: Nature, v. 521, no. 7553, p. 436–444, doi:10.1038/nature14539.

Lin, LI 1989, A concordance correlation coefficient to evaluate reproducibility: Biometrics, v. 45, no. 1, p. 255–268, doi:10.2307/2532051.

Maas, AL, Hannun, AY and Ng, AY 2013, Rectifier nonlinearities improve neural network acoustic models, *in* Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 16–21 June 2013, <https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf>.

Ormsby, WR, Thom, J, Howard, SHD, Then, D, Gardiner, N and Tapping, B 2021, Mean soil sample geochemical data, *in* Accelerated Geoscience Program extended abstracts *compiled by* Geological Survey of Western Australia: Geological Survey of Western Australia, Record 2021/4, p. 32–33.

Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, Vanderplas, J, Passos, A, Cournapeau, D, Brucher, M, Perrot, M and Duchesnay, E 2011, Scikit-learn: machine learning in Python: Journal of Machine Learning Research, v. 12, p. 2825–2830, doi:10.5555/1953048.2078195.

Puzyrev, V 2019, Deep learning electromagnetic inversion with convolutional neural networks: Geophysical Journal International, v. 218, no. 2, p. 817–832, doi:10.1093/gji/ggz204.

Puzyrev, V, Zelic, M and Duuring, P 2023, Applying neural networks-based modelling to the prediction of mineralization: a case-study using the Western Australian Geochemistry (WACHEM) database: Ore Geology Reviews, v. 152, p. 105242, doi:10.1016/j.oregeorev.2022.105242.

Savitzky, A and Golay, MJE 1964, Smoothing and differentiation of data by simplified least squares procedures: Analytical Chemistry, v. 36, p. 1627–1639, doi:10.1021/ac60214a047.

Srivastava, N, Hinton, G, Krizhevsky, A, Sutskever, I and Salakhutdinov, R 2014, Dropout: a simple way to prevent neural networks from overfitting: Journal of Machine Learning Research, v. 15, p. 1929–1958.

Tofallis, C 2015, A better measure of relative prediction accuracy for model selection and model estimation: Journal of the Operational Research Society, v. 66, no. 8, p. 1352–1362, doi:10.1057/jors.2014.103.

Virtanen, P, Gommers, R, Oliphant, TE, Haberland, M, Reddy, T, Cournapeau, D, Burovski, E, Peterson, P, Weckesser, W, Bright, J, van der Walt, SJ, Brett, M, Wilson, J, Millman, KJ, Mayorov, N, Nelson, ARJ, Jones, E, Kern, R, Larson, E, Carey, CJ, Polat, İ, Feng, Y, Moore, EW, VanderPlas, J, Laxalde, D, Perktold, J, Cimrman, R, Henriksen, I, Quintero, EA, Harris, CR, Archibald, AM, Ribeiro, AH, Pedregosa, F and van Mulbregt, P 2020, SciPy 1.0: fundamental algorithms for scientific computing in Python: Nature Methods, v. 17, no. 3, p. 261–272, doi:10.1038/s41592-019-0686-2.

Xu, B, Wang, N, Chen, T and Li, M 2015, Empirical evaluation of rectified activations in convolutional network: arXiv.org, doi:10.48550/arXiv.1505.00853.

Appendix 1.

Error codes

The following WAMEX entries were treated as unconventional error codes and replaced with Not a Number (NaN) value.

Conventional codes:

−9999, −6666.

Non–conventional codes (found in surface datasets):

−999, −99999, −9999000, −9990000, −99990000, −999000, −99000, −99900, −990000, −9989990, −9005000, −9000000, −99990000000, −99999000, −9960000, −9970000, −9009, −999.9, −99.99, −5555, −4444, −7777, −5559, −5559000, −5555000, −55550000, −5557, −5557000, −5666, −5666000, −5.556, −55590000, −5556, −5556000, −5.5550003, −4440000, −4444000, −7777000, −10000000, −100000000, −20000000, −666, −66660000, −6666000, −6660000, −666000, −3333, −3333000, −33300000, −33330000, −33330000000, −60000000, −1e32, −50000000000, −9910000, −55570000, −55560000, −5560000, −555, −555000, −99000000, −990000000, −8880000, −888000, −888, −8888, 20000000, 40000000.

Appendix 2.

Oxide conversion factors

Field name	*Ratio*	*Field name*	*Ratio*
Fe_2O_{3T}_pct	7776.0	BaO_ppm	0.89566
Fe_ppm	1286.0	Ce_ppm	1.2284
Al_ppm	0.000188946265571534	CeO_2_ppm	0.814089
Al_2O_3_pct	5292.51	Cr_ppm	1.46155667478318
Ca_ppm	0.000139918651296136	Cr_2O_3_ppm	0.684202
CaO_pct	7147.01	Cs_ppm	1.0602
K_ppm	0.000120460593123868	Cs_2O_ppm	0.943226
K_2O_pct	8301.47	Dy_ppm	1.1477
Mg_ppm	0.000165827579116338	Dy_2O_3_ppm	0.871318
MgO_pct	6030.36	Er_ppm	1.1435
Mn_ppm	0.000129122727278597	Er_2O_3_ppm	0.87452
MnO_pct	7744.57	Eu_ppm	1.1579
Na_ppm	0.000134796867859978	Eu_2O_3_ppm	0.86361
Na_2O_pct	7418.57	Ga_ppm	1.3442
P_ppm	0.000229136544758387	Ga_2O_3_ppm	0.743925
P_2O_5_pct	4364.21	Gd_ppm	1.1526
Si_ppm	0.000213931657392729	Gd_2O_3_ppm	0.867591
SiO_2_pct	4674.39	Ho_ppm	1.1455
Ti_ppm	0.00016680344549197	Ho_2O_3_ppm	0.872973
TiO_2_pct	5995.08	La_ppm	1.1728
Ba_ppm	1.11649509858652	La_2O_3_ppm	0.85268
Li_ppm	2.1527	Tb_ppm	1.151
Li_2O_ppm	0.46457	Tb_2O_3_ppm	0.868803
Lu_ppm	1.1371	Th_ppm	1.1379
Lu_2O_3_ppm	0.879383	ThO_2_ppm	0.878809
Nb_ppm	1.4305	Tm_ppm	1.1421
Nb_2O_5_ppm	0.699044	Tm_2O_3_ppm	0.875609
Nd_ppm	1.1664	U_ppm	1.17924250178655
Nd_2O_3_ppm	0.857351	U_3O_8_ppm	0.848002
Pr_ppm	1.1703	V_ppm	1.78518510584362
Pr_2O_3_ppm	0.854469	V_2O_5_ppm	0.560166
Rb_ppm	1.0936	W_ppm	1.261
Rb_2O_ppm	0.914412	WO_3_ppm	0.793
Sb_ppm	1.3284	Y_ppm	1.2699
Sb_2O_5_ppm	0.83534	Y_2O_3_ppm	0.78744
Sm_ppm	1.1596	Yb_ppm	1.1387
Sm_2O_3_ppm	0.86239	Yb_2O_3_ppm	0.878201
S_pct	2.49713453811751	Zn_ppm	1.2448
SO_3_pct	0.400459	ZnO_ppm	0.803397
Sr_ppm	1.18259923485829	Zr_ppm	1.350787306381520
Ta_ppm	1.2211	ZrO_2_ppm	0.740309
Ta_2O_5_ppm	0.818967		

Appendix 3.

Analyte detection limits

	WACHEM			*ALS*			*ActLabs*			*Final limit*	*Unit (if not ppm)*
Element	Method	Unit	Limit	Method	Unit	Limit	Method	Unit	Limit		
Ag	ME-MS61	ppm	0.01	AuME-ST43	ppm	0.001				0.001	
Al_2O_3	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Al				AuME-ST43	%	0.01				10	
As	ME-MS42	ppm	0.1	ME-MS23	ppm	0.0005				0.0005	
Au	PGM-ICP24	ppm	0.001	Au-CN43	ppm	0.00002				0.00002	
B				AuME-ST43	ppm	2	Ultratrace 1	ppm	1	1	
Ba	ME-MS81	ppm	0.5	ME-MS23	ppm	0.01				0.01	
Be				ME-MS23	ppm	0.002				0.002	
BaO	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Bi	ME-MS42	ppm	0.01	ME-MS23	ppm	0.0003				0.0003	
Br				ME-HAL01	ppm	0.02				0.02	
C	C-IR07	%	0.01	ME-IR08	%	0.01				100	
CaO	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Ca				AuME-ST43	%	0.01				5	
Cd	ME-MS61	ppm	0.02	ME-MS23	ppm	0.0002				0.0002	
Ce	ME-MS81	ppm	0.1	ME-MS23	ppm	0.0001				0.0001	
Cl				ME-HAL01	ppm	0.1				0.1	
Co	ME-MS61	ppm	0.1	ME-MS23	ppm	0.0003				0.0003	
Cr	ME-MS81	ppm	10	ME-MS23	ppm	0.001				0.001	
Cr_2O_3	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Cs	ME-MS81	ppm	0.01	ME-MS23	ppm	0.0001				0.0001	
Cu	ME-MS61	ppm	0.2	ME-MS23	ppm	0.001				0.001	
Dy	ME-MS81	ppm	0.05	ME-MS23	ppm	0.0001				0.0001	
Er	ME-MS81	ppm	0.03	ME-MS23	ppm	0.0001				0.0001	
Eu	ME-MS81	ppm	0.03	ME-MS23	ppm	0.0001				0.0001	
F				ME-HAL01	ppm	0.05				0.05	
Fe_2O_{3T}	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Fe				ME-MS23	%	0.00001				0.1	
Ga	ME-MS81	ppm	0.1	ME-MS23	ppm	0.0005				0.0005	
Ge				ME-MS23	ppm	0.0001				0.0001	
Gd	ME-MS81	ppm	0.05	ME-MS23	ppm	0.0001				0.0001	
Hf	ME-MS81	ppm	0.2	ME-MS23	ppm	0.00005				0.00005	
Hg	ME-MS42	ppm	0.005	ME-MS23	ppm	0.0001				0.0001	
Ho	ME-MS81	ppm	0.01	ME-MS23	ppm	0.0001				0.0001	
I				ME-HAL01	ppm	0.002				0.002	
In	ME-MS42	ppm	0.005	ME-MS23	ppm	0.0001				0.0001	
Ir				PGM-MS25NS	ppm	0.001	1B1	ppm	0.0001	0.1	ppb
K_2O	ME-XRF26	%	0.01	ME_XRF26	%	0.01				0.01	%
K				AuME-ST43	%	0.01				5	
La	ME-MS81	ppm	0.1	ME-MS23	ppm	0.0001				0.0001	
Li	ME-MS61	ppm	0.2	ME-MS23	ppm	0.0002				0.0002	
LOI	ME-GRA05	%	0.01	OA-GRA05	%	0.01				0.01	%
Lu	ME-MS81	ppm	0.01	ME-MS23	ppm	0.0001				0.0001	
MgO	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Mg				ME-MS23	%	0.000001				0.01	
MnO	ME-XRF26	%	0.01	ME_XRF26	%	0.01	WRA-ICP-4B	%	0.001	10	

	WACHEM			*ALS*			*ActLabs*			*Final limit*	*Unit (if not ppm)*
Element	Method	Unit	Limit	Method	Unit	Limit	Method	Unit	Limit		
Mn				ME-MS23	ppm	0.01				0.01	
Mo	ME-MS61	ppm	0.05	ME-MS23	ppm	0.0005				0.0005	
Na_2O	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Na				AuME-ST43	%	0.001				2	
Nb	ME-MS81	ppm	0.2	ME-MS23	ppm	0.0001				0.0001	
Nd	ME-MS81	ppm	0.1	ME-MS23	ppm	0.0001				0.0001	
Ni	ME-MS61	ppm	0.2	ME-MS23	ppm	0.001				0.001	
Os				PGM-MS25NS	ppm	0.002				2	ppb
P_2O_5	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
P				AuME-ST43	%	0.0005				5	
Pb	ME-MS61	ppm	0.5	ME-MS23	ppm	0.0001				0.0001	
Pd	PGM-ICP24	ppm	0.001	ME-MS23	ppm	0.00005				0.05	ppb
Pr	ME-MS81	ppm	0.03	ME-MS23	ppm	0.0001				0.0001	
Pt	PGM-ICP24	ppm	0.001	PGM-MS23L	ppm	0.0001				0.1	ppb
Rb	ME-MS81	ppm	0.2	ME-MS23	ppm	0.0001				0.0001	
Re	ME-MS42	ppm	0.001	ME-MS23	ppm	0.00001				0.00001	
Rh				Rh-MS25	ppm	0.001	1B1	ppm	0.0002	0.2	ppb
Ru				PGM-MS25NS	ppm	0.003				3	ppb
S	S-IR08	%	0.01	AuME-ST43	%	0.002				0.002	%
Sb	ME-MS42	ppm	0.05	ME-MS23	ppm	0.00005				0.00005	
Sc	ME-MS61	ppm	0.1	ME-MS23	ppm	0.001				0.001	
Sc-2	ME-MS42	ppm	0.1								
Se	ME-MS42	ppm	0.2	AuME-ST43	ppm	0.002				0.002	
SiO_2	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Si										10	
Sm	ME-MS81	ppm	0.03	ME-MS23	ppm	0.0001				0.0001	
Sn	ME-MS81	ppm	1	ME-MS23	ppm	0.0002				0.0002	
SO_3	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Sr	ME-MS81	ppm	0.1	ME-MS23	ppm	0.001				0.001	
SrO	ME-XRF26	%	0.01	ME_XRF26	%	0.01				100	
Ta	ME-MS81	ppm	0.1	ME-MS23	ppm	0.00005				0.00005	
Tb	ME-MS81	ppm	0.01	ME-MS23	ppm	0.0001				0.0001	
Te	ME-MS42	ppm	0.01	ME-MS23	ppm	0.0005				0.0005	
Th	ME-MS81	ppm	0.05	ME-MS23	ppm	0.00002				0.00002	
TiO_2	ME-XRF26	%	0.01	ME_XRF26	%	0.01	WRA-ICP-4B	%	0.001	10	
Ti				ME-MS23	%	0.0000005				0.005	
Tl	ME-MS42	ppm	0.02	ME-MS23	ppm	0.00005				0.00005	
Tm	ME-MS81	ppm	0.01	ME-MS23	ppm	0.0001				0.0001	
U	ME-MS81	ppm	0.05	ME-MS23	ppm	0.00005				0.00005	
V	ME-MS81	ppm	5	ME-MS23	ppm	0.0002				0.0002	
W	ME-MS81	ppm	1	ME-MS23	ppm	0.0001				0.0001	
Y	ME-MS81	ppm	0.1	ME-MS23	ppm	0.0001				0.0001	
Yb	ME-MS81	ppm	0.03	ME-MS23	ppm	0.0001				0.0001	
Zn	ME-MS61	ppm	2	ME-MS23	ppm	0.01				0.01	
Zr	ME-MS81	ppm	2	ME-MS23	ppm	0.0001				0.0001	

REPORT
237

DEEP-LEARNING IDENTIFICATION OF ANOMALOUS DATA IN GEOCHEMICAL DATASETS

V Puzyrev, P Duuring, SHD Howard, JR Lowrey, WR Ormsby,
D Purnomo, D Then and J Thom

The Western Australian Mineral Exploration (WAMEX) database contains geochemical data provided to the Geological Survey of Western Australia (GSWA) in digital format by the exploration and mining industry. The WAMEX database is known to contain a significant amount of spurious data, including errors in unit reporting and incorrect assignment of analytes brought about mainly by errors in post-analysis data reporting and, in some cases, due to low accuracy of the analytical technique. There are significant time and cost challenges in manually identifying and correcting these issues.

In this study, a set of deep-learning methods was applied to the harmonized surface and drillhole WAMEX datasets to identify (and replace) potential spurious data and estimate missing analyte values wherever possible. The method was entirely data-driven and, after the corresponding networks have been trained, allows the results to be obtained instantly. Deep-learning methods delivered good results at modest computational cost and, contrary to many other statistical methods, required no manual feature engineering. The results of this study demonstrate the efficacy of the method for the different types of geochemical data included in the WAMEX database (i.e. surface vs drillhole sample media, and different laboratory analytical methods).

Further details of geoscience products are available from:

First Floor Counter
Department of Mines, Industry Regulation and Safety
100 Plain Street
EAST PERTH WESTERN AUSTRALIA 6004
Phone: +61 8 9222 3459   Email: publications@dmirs.wa.gov.au
www.dmirs.wa.gov.au/GSWApublications