The following document is a pre-print version of:

Schnitzler N, Ross P-S, Gloaguen E (2019) Using machine learning to estimate a key missing geochemical variable in mining exploration: application of the Random Forest algorithm to multi-sensor core logging data. J. Geochem. Explor. 205:article 106344

Using machine learning to estimate a key missing geochemical variable in mining exploration: application of the Random Forest algorithm to multi-sensor core logging data

N. Schnitzler[1,2], P.-S. Ross[1,*], E. Gloaguen[1]

1. Institut national de la recherche scientifique, 490 de la Couronne, Québec (Qc), G1K 9A9, Canada

2. Now at: Sirios Resources Inc., 1000 St-Antoine Ouest, suite 410, Montréal (Qc), H3C 3R7, Canada

*Corresponding author, rossps@ete.inrs.ca

Abstract

Mining exploration increasingly relies on large, multivariate databases storing data ranging from drill core geochemical analysis to geophysical data or geological descriptions. Utilizing these large datasets to their full potential implies the use of multivariate statistical analysis such as machine learning. The Random Forest algorithm has proved its efficiency in mining applications. In this study we use it to estimate a key geochemical element, sodium, using a multivariate chemo-physical dataset measured on drill cores in the Matagami mining district of Québec, Canada. Sodium is important to characterize hydrothermal alteration in volcanogenic massive sulfide settings, since Na depletion can be used to vector towards ore, but this element is not readily measured by portable X-ray fluorescence (pXRF). We first test the algorithm on a database of over 8000 traditional laboratory geochemistry analyses and find a correlation of 0.95 between estimated and measured Na. We then test the algorithm on the multi-sensor core logging data, including density, magnetic susceptibility, and 15 geochemical elements by pXRF, but borrowing Na from traditional geochemistry (n = 260). This yields correlations of 0.66 to 0.75 depending on the training and testing sets. Finally the algorithm is applied to the whole multiparameter database (n = 9675) to estimate Na downcore. There is a good general correspondence with the downcore Na patterns seen through traditional geochemistry, and the estimated Na which has much greater spatial resolution. Random Forest appears to be a very good estimation tool when using large amounts of data and variables, as it uses all variables and automatically prioritizes the most useful. This method also allows visualization of the weight of each variable in the estimation. Future studies should compare RF with other methods.

1. Introduction

Modern mining exploration increasingly relies on processing large, multivariate databases which includes data coming from various sources (drill core analysis, geophysics, geological mapping, …) and ranges from qualitative to quantitative (e.g. the recent Integra Gold Rush competition). Even at the diamond drill core characterization stage, which used to consist primarily of a visual log by the geologist, more and more data is becoming available (e.g., physical rock properties, geochemistry, mineralogy, …) (e.g., Ross et al., 2013, 2016a; Jácomo et al., 2015; Ross and Bourke, 2017; Wang et al., 2017; Bérubé et al., 2018; Chen et al., 2018). Utilizing such large multi-parameter datasets to their full potential requires specific algorithms such as multivariate statistical analysis (e.g., Fresia et al., 2017) or ensemble trees (e.g., Bérubé et al., 2018; Caté et al., 2018; Chen et al., 2018). Artificial intelligence methods are already used by some mining companies, but generally remain little known in the mining sector. However they have proven themselves in many other applied fields (diagnostic systems in hospitals, electrical network management, image processing, ...). A particularly relevant branch of artificial intelligence for the mining sector is supervised machine learning (Heutte et al., 2008; Caté et al., 2017). Supervised learning brings together machine learning techniques that automatically produce rules from a learning database that contains "examples". Once the rule is generated and tested, it is applied to a set of inputs (e.g., variables measured along the drill core) to predict outputs (e.g., gold grades or any other sought-after parameter) (Mohamadally and Fomani, 2006; Fischer, 2014).

In this article, we apply a supervised machine learning method, the Random Forest (RF) algorithm, to a multi-sensor drill core logging database. This algorithm,

developed by Breiman in 1984 (see Breiman, 2001), can be used for classification (predicting a categorical variable such as lithology) or regression (predicting a continuous variable such as a geochemical element). Here we focus on the latter application. This study was conducted at the McLeod volcanogenic massive sulphide (VMS) deposit in the Matagami mining camp, Quebec, Canada. We use the RF algorithm to estimate a missing parameter, sodium, from multiparameter data which contains density, magnetic susceptibility, 15 geochemical elements, average visible light reflectance, and infrared spectrometry. Sodium is missing or difficult to measure by portable x-ray fluorescence (pXRF) analyzers, a commonly employed technology to acquire in situ geochemical measurements on drill cores (e.g., Ross et al., 2014a, 2014b). Yet sodium variations can be used to characterize the hydrothermal alteration that rocks have undergone around a VMS deposit (e.g., Large et al., 2001; Franklin et al., 2005; Gifkins et al., 2005), and are therefore very useful for mining exploration. We briefly present the data acquisition methods, then all the tests performed and the results obtained for the estimation of sodium with the RF algorithm. A first step validates the method using known geochemical data from our industrial partner, then the same method is applied on the multiparameter data. We conclude that the RF algorithm is a good tool in this case.

2. Geological context

The McLeod VMS deposit, around which this study was made, is located in the Matagami mining district, in the northern part of Abitibi Subprovince, Québec (Fig. 1a). VMS deposits of the Matagami district are mainly found along two bands, the North Flank and the South Flank (Fig. 1b). An area further west, informally referred to as the "West Camp", forms a third prospective band for exploration. The McLeod deposit – the topic of this investigation – and the Bracemac deposit, 1 km to the NW, are currently exploited together by Glencore as part of the Bracemac-McLeod mine (Fig. 2).

The lithological succession at McLeod is typical of the South Flank. It dips about 70° towards the SW, its stratigraphic summit being in the same direction (Fig. 3). The succession includes the Watson Lake and Wabassee groups (Sharpe, 1968; Debreil et al., 2018). Given the great thickness of the Wabassee Group, the lithologic succession is dominated by mafic to intermediate flows with a smaller proportion of felsic flows. Mineralization in the South Flank is mainly found at the level of the Key Tuffite (e.g., Adair, 2009; Genna et al., 2014), the main marker horizon in the South Flank. At McLeod, the Key Tuffite and the ore zones are sandwiched between the Watson Lake Rhyolite (footwall) and the Bracemac Rhyolite (immediate hangingwall). The Bracemac Rhyolite is overlain by the Bracemac Tuffite, then by mafic to intermediate rocks of the Wabassee Group (Fig. 3). Volcanic rocks are cut by intrusions of gabbro and diorite and rare felsic intrusions.

Ore lenses are composed mainly of pyrite, chalcopyrite, sphalerite, and sometimes pyrrhotite. Many VMS deposits in the South Flank are surrounded by discordant hydrothermal alteration, but at McLeod, alteration is semi-concordant with stratigraphy (Fig. 3). The proximal and intense alteration underlies the mineralized lenses, at the top of the Watson Lake Rhyolite. This zone corresponds to intense chloritization with a leaching of silica, sodium, calcium and potassium. In the most altered rocks, Na concentration falls close to zero. In some drillholes, there is also intense alteration just above the mineralized lenses in the lower part of the Bracemac Rhyolite. Moving laterally away from the mineralization, chlorite alteration becomes less intense (Genna et al., 2014), sericite increases, and Na increases, both in the footwall and in the hanging wall.

3. Methods

3.1 Datasets

The data used here was acquired on exploration drill cores, mostly of NQ caliber, and consists of two types. The first type of data is traditional, destructive, laboratory lithogeochemistry on 10-25 cm-long whole core samples, with a downcore sample spacing on the order of 30 m, available for 314 drillholes from the entire Bracemac-McLeod area, courtesy of Glencore (n = 8287 analyses). This first dataset therefore includes the full suite of rock types (except ore zones) and Na concentrations for the whole area. This dataset covers a larger geographic area than the second one (as detailed below), but the geology is the same at Bracemac and McLeod.

The second dataset consists of non-destructive multi-sensor core logging data acquired by *Institut national de la recherche scientifique* on nine selected drillholes located 0-1.5 km ESE of the McLeod deposit (Ross et al., 2016b). After a detailed visual relogging of the nine drillholes by the first author, the multi-sensor dataset was acquired at the mine site using a mobile laboratory called the *Laboratoire mobile de caractérisation physique, minéralogique et chimique des roches* (LAMROC). The laboratory measures magnetic susceptibility, density with gamma ray attenuation, alteration mineralogy with infrared spectrometry, and geochemistry through pXRF (15 elements). The downcore sample spacing was about 30 cm (n = 9675 data points). The measured data set was corrected (to fix calibration issues) and compiled as downhole profiles. An example of these profiles is given here for drillhole MCL-12-09 (Figs. 4 to 7) and profiles for the eight other holes are available in Ross et al. (2016b) and Schnitzler (2017). For further methodological details, the reader is referred to Ross et al. (2013, 2014a, 2014b), Fresia (2013), Bourke and Ross (2016) and Ross and Bourke (2017).

The large amount of data and the diversity of parameters (several thousand measurement points per drillhole with more than 20 variables) generated by LAMROC increases the spatial resolution and quantity of information, which in turn allows a better understanding of the geology around a

deposit. This abundance of collocated measures opens the way for quantitative integration and the prediction of exploration vectors. On the other hand, this requires new approaches to quantitative assimilation of data to facilitate interpretation.

3.2 Data exploration

Before applying any algorithm, a data exploration step is required. We computed basic statistics, histograms, cross-plots and a correlation matrix using the first dataset, i.e. whole-rock geochemistry from Glencore (n = 8287 analyses). We did this with the unprocessed geochemical variables and then again with the major oxides transformed into centered log-ratio (CLR) values. The cross-plots and histograms (available upon request from the corresponding author) show that the variables do not display a Gaussian or even symmetric distribution, and that the variables have complex relationships. Furthermore, the correlations between the primary variable (here Na_2O) and the other variables are weak: the maximum correlation is -0.7 in the CLR-transformed data (Fig. 8), and only 0.4 in the untransformed data (correlation matrix for raw data available upon request from the corresponding author).

In summary, the correlations are weak, the data do not show Gaussian distributions, and the data are clustered in multiple classes. This prevents the use of methods relying on strict stationarity like multiple regression or support vector machine. Therefore we selected the random forest algorithm, which does not have this limitation.

3.3 The Random Forest algorithm

The machine learning method selected here, which allows the estimation of sodium or any other missing parameter from a dataset, is the Random Forest (RF) algorithm. This choice is justified by the simplicity of use, its versatility and its ability to quantitatively rank the most important variables in the training step (Rakotomalala, 2005; Heutte et al., 2008; Carranza and Laborte, 2015; Rodriguez-Galiano et al., 2015). The RF method was utilized successful by a team that included the first author during the International Integra Gold Rush competition (2015), which was aimed at predicting targets for gold exploration.

Random Forest is part of a family referred to as "bagging" algorithms, which refers to bootstrap averaging methods. They allow the user to build optimal decision trees based on the aggregation of multiple iterative trees built from randomly selected samples of the training step. The aggregation and the construction of small trees from small subset allows for an increase of the prediction on noisy data sets. The training set consists of a part of the database for which all the variables are known, including the one that we want to estimate later (here, sodium). This training step permits to calibrate the parameters that optimize the prediction of sodium (number of trees, leaf size, size of the random subset, …). The algorithm is then applied to a test subset for which the parameter to be estimated has been set aside for validation purposes. Finally, if these two steps are satisfactory, the prediction step can be deployed.

3.4 Implementation

For this project, sodium concentrations were estimated using an RF regression algorithm (Python Code RandomForestRegressor, from Pedregosa et al., 2011). Several test and validation steps were performed to validate the method (Table 1; Schnitzler, 2017). The first three steps involved the use of Glencore's lithogeochemical data – which *does* include Na – to check the performance of the algorithm. The fourth step allowed us to explore the use of variables coming from the LAMROC multiparameter database. Finally, the fifth and final step was the prediction of sodium from LAMROC measurements. Note that all of these steps are independent of each other: each time, the algorithm was re-run with the specific data and variables listed in Table 1 and described below. In other words, what the machine learned in a specific step did not influence what went on in subsequent steps.

Step 1. In more detail, our first step was to validate the chosen method using Glencore's conventional lithogeochemistry data. Only the chemical elements that were systematically analyzed in all drill cores were retained as variables. This represents 18 elements and oxides: Al_2O_3, Ba, CaO, Cr_2O_3, Fe_2O_3, K_2O, LOI (loss on ignition, the loss of mass resulting from heating a powdered material to a very high temperature), MgO, MnO, Na_2O, Nb, P_2O_5, Rb, SiO_2, Sr, TiO_2, Y and Zr. Measurements below the detection limit were replaced by half the value of the detection limit for each element, to avoid zeros. This choice avoids modifying the structure of the data and limits the loss of information (Martín-Fernández et al., 2003). The algorithm was applied with the unprocessed geochemical variables and then again with the major oxides transformed into CLR values. The training set corresponded to 2/3 of the measurements, taken randomly ($n_{training}$ = 5524 data points). The algorithm was then applied to the test subset (the remaining 1/3 of the data, n_{test} = 2763) for which sodium had been set aside for validation purposes.

Step 2. The same process was performed on the Glencore geochemical data taken only on the nine drillholes selected for the LAMROC multiparameter study, in order to verify the results when the number of data decreased (n = 260). The variables used initially were the same 18 geochemical elements and oxides as those used in step 1. We also added two other variables, Cu and Zn, to verify the impact ($n_{variable}$ = 20). The data were randomly divided 2/3 – 1/3 to form the training set ($n_{training}$ = 173) and the test set (n_{test} = 87).

Step 3. In the third step, the RF algorithm was tested on three different training and test sets (A, B and C), keeping the same variables as before ($n_{variable}$ = 20). These tests were done without and with the CLR transformation on the Glencore geochemical data for the nine drill holes.

Step 4. To train and test the algorithm on LAMROC data, it was necessary to know the sodium concentration at certain points, but of course the LAMROC data does not contain it. Therefore Na for each traditional geochemical analysis for the nine holes was "loaned" to the nearest LAMROC data point. It was not possible to have perfectly colocated data since the traditional geochemistry, done in this case before pXRF analysis, had destroyed the sample. Even though pXRF and conventional geochemistry do not involve measurements on the same volume of rock, we did not explicitly consider the different support sizes. A "change of support analysis" is not needed for machine learning algorithms. In conventional geostatistics and regressions, links between variables are linear and variances are important. However this does not apply to ensemble methods (Rashka and Mirjalili, 2017).

This gave a total of 260 measurements (LAMROC data plus borrowed Na) for the nine boreholes studied. This was randomly divided into a training set corresponding to 2/3 of the measurements ($n_{training}$ = 173) and a test set for which Na was set aside (n_{test} = 87). The same three training and test sets as before were used (i.e., A, B and C). For each set, the algorithm was first incremented with all variables measured by LAMROC ($n_{variable}$ = 20), including geochemistry (Al_2O_3, CaO, Cr, $Fe_2O_3^t$, K_2O, MgO, MnO, Nb, Ni, Rb, SiO_2, Sr, TiO_2, Y, Zr), average visible light reflectance, physical properties (density, magnetic susceptibility) and near-infrared mineralogy expressed as two mineral groups per sample (e.g., Fig. 7, black diamonds and red squares). These "group 1" and "group 2" mineralogical variables were used because they did not require any treatment. Then the algorithm was used with all the variables, but transforming the major oxides (including Na_2O) into CLR ($n_{variable}$ = 20) and finally removing the chemical elements having more than 10% of data points below the limit of detection ($n_{variable}$ = 13).

Step 5. In the final step, the algorithm was used with all the variables measured by LAMROC, on the unprocessed data ($n_{variable}$ = 20). The training set consisted of the 260 LAMROC measurement points with the borrowed Na. The algorithm was applied to the prediction set corresponding to all measurements made by LAMROC (n = 9675 measurements), so that Na could be estimated everywhere along the nine drill holes.

4. Results

4.1 Step 1

Step 1, testing the RF algorithm on a large traditional geochemistry database for the Bracemac-McLeod area, shows a Pearson correlation of 0.93 on unprocessed data (Table 1, Fig. 9a) and 0.95 with the CLR transformation for the major oxides (Table 1, Figs. 9c). The RF method allows an excellent estimation of Na in this test, regardless of whether the unprocessed data or the CLR transformed data is used.

However, the CLR transformation has a large impact on which variables contribute the most to the estimation. When the variables are not transformed, Sr is the most influential element (35%), then iron (13%), while MgO is only the 10th most important variable (Fig. 9b). Sr is an element contained in particular in plagioclase, which is likely to be destroyed during alteration in VMS settings (Franklin et al., 2005). With the CLR transformation, Mg is clearly the most influential variable (57%), followed by iron (10%) (Fig. 9d). The link between Na and Fe-Mg might be related to the chloritization of rhyolites (gains in Fe-Mg and leaching of Na; Piché, 1991), even if there is no simple linear relationship between Fe-Mg and Na.

4.2 Step 2

In the second step, with CLR-transformed traditional geochemistry for only 9 drill holes, and the same 18 variables, the results show a correlation of 0.84 (Fig. 10a). Here, LOI and Sr are the most influential variables (31% and 25% respectively; Fig. 10b). LOI generally increases with hydrothermal alteration in greenschist facies rocks, and so might be correlated to Na leaching. Adding copper and zinc ($n_{variable}$ = 20) yields a nearly identical correlation of 0.83 (Fig. 10c), and LOI and Sr remain the most influential variables (30% and 24% respectively, Fig. 10d). Overall, this second step shows that using a much smaller dataset reduces the quality of the estimation, but the correlation is still acceptable, and most data points are correctly estimated.

4.3 Step 3

The third step explores the effect varying the training and test sets (A, B and C), without and with the CLR transformation. Overall, the correlation coefficients vary from 0.82 to 0.88 (Table 1, Fig. 11). The CLR transformation improves the correlation in two cases out of three (random sets A and B, Fig. 11). In step 3, Sr is the most, or second most, influential variable (Fig. 12), and Fe is in the top three, but other variables vary in importance from one random set to another. This indicates a fairly high level of noise in the database.

4.4 Step 4

LAMROC data is used in step 4, with the borrowed Na, without and with the CLR transformation, on 13 to 20 variables and three training and test sets (Table 1). Of the three tests listed in Table 1 for step 4, only two are illustrated in figure 13: 20 versus 13 variables, without the CLR transformation. The correlation coefficients vary from 0.66 to 0.75 overall (Table 1). The CLR transformation of major oxides and the suppression of elements with more than 10% of data below the detection limit ($n_{variable}$ = 13) do not systematically increase the correlation.

The most influential variable is Sr (30-40%), generally followed by density (10-18%) (Table 1, Fig. 14). We note that the two physical properties, density and magnetic susceptibility, are typically among the ten most important

variables, and sometimes among the five most important, for estimating Na.

Step 4 shows that with almost similar data points (n_{total} = 260, on the same nine drillholes), the use of LAMROC data instead of traditional geochemistry decreases the quality of Na estimation. Possible reasons for this include: (1) during step 4, Na has to be borrowed from traditional geochemistry on an adjacent sample, i.e. the loaned Na and the other variables entering the RF algorithm are not perfectly co-located; (2) pXRF data is not as accurate and precise as traditional geochemistry. But the estimated Na should still be useful to show trends.

4.5 Step 5

The final step is the estimation of Na for all LAMROC data points. We show estimated Na as a function of depth for four drillholes, placed from proximal to distal relative to the McLeod VMS deposit (Fig. 15); the profiles for the other five holes are available in Schnitzler (2017). The graphs show a good general correspondence in Na measured by traditional geochemistry versus its estimate from the LAMROC data. Although the Na estimation is far from perfect, the advantage of the large amount of data provided by LAMROC is to be able to observe trends where there is no traditional geochemistry information, which is over 99% of the length of the drill core, assuming the average sample length is 20 cm and the sampling interval is 30 m, for traditional geochemistry.

5. Discussion

5.1 Effectiveness of the RF algorithm to estimate a key missing variable

Step 1 of our methodology shows that a RF algorithm, applied to a very large database of traditional geochemistry (n_{data} > 8000), is excellent at predicting Na. In steps 2 and 3, we drastically cut the number of traditional geochemistry samples by using only the nine drill holes selected for this project (n_{data} = 260). We explored the influence of adding extra variables, of doing the CLR transformation on major oxides or not, and of different training and test sets. Correlations for steps 2 and 3 ranged from 0.82 to 0.88, i.e. lower than in step 1, but still satisfactory. In these tests, adding Cu and Zn to the geochemical dataset yielded no benefit, whereas the impact of the CLR transformation was sometimes positive, and sometimes not. Changing the training and test sets was the main factor generating variability in the correlations coefficients, showing that the RF algorithm is less robust for these much smaller training sets than it was in step 1.

In step 1, with unprocessed data, the most important variables were Sr, Fe_2O_3 and SiO_2; with the CLR transformation, it was MgO, Fe_2O_3, and LOI. In steps 2 and 3, LOI or Sr became the most influential variables, or even Fe_2O_3 in some cases, depending on the training and test sets. The variability in the importance of the variables may be due to outliers in the training set or in the test set.

After these three steps using traditional geochemistry from Glencore, we were confident that the RF algorithm would allow Na estimation when applied to the multiparameter LAMROC data. In step 4, we paired the LAMROC data with Na borrowed from traditional geochemistry. We tested the impact of the CLR transformation, of removing chemical elements having more than 10% of their measurements below the detection limit, and of changing the training and test sets. Correlations were in the range 0.66-0.75, and changing the training and test sets was again the most important factor influencing the correlation coefficient. This variability may be due to outliers. It is interesting to note that in step 4, the second most important variable was typically density, which indicates that the addition of physical properties can be helpful to estimate a geochemical variable.

Step 4 showed that the RF algorithm could be used on the LAMROC data, so in step 5 we estimated Na for the whole dataset (n_{data} = 9625), using the non-transformed pXRF data, two physical properties, average visible light reflectance, and some mineralogical parameters. We obtained estimated downhole Na profiles at high spatial resolution for the nine studied holes. Comparison with analytically accurate but low spatial resolution traditional geochemistry shows a reasonable overall agreement, although there are some discrepancies (Fig. 15). The estimation would have been better had a larger training set been available. Despite the imperfect correlation between measured Na and estimated Na, the spatial resolution of the estimated Na is two orders of magnitude better.

5.2 Na variations with distance from ore

Along the studied drill holes that pass through the sulphide lenses, the Na concentration gradually decreases to near zero, about 60 m above the mineralization, and stays low for over 100 m below the ore zone (Fig. 15, MCL-12-09). This is the mineralization-proximity signal related to hydrothermal alteration in the footwall and hangingwall rhyolites (the Watson Lake Rhyolite and Bracemac Rhyolite, respectively). This signal decreases in strength laterally away from the ore zone. Some 50 m further, in MC-04-11, the loss of sodium is mostly localized in the Watson Lake Rhyolite (Fig. 15). Finally, there is no Na depletion pattern in MC-05-19/19A and MC-09-76, respectively 200 and 600 m laterally from the sulphide lens.

5.3 Other possible uses of machine learning in mining exploration

Our approach illustrates how a large multiparameter database (here 9675 measurement points, 20 parameters) can be exploited and enhanced using machine learning methods. This case study demonstrates how to estimate parameters that are not directly measured, but may be important for mineral exploration. Here we have estimated Na to assess hydrothermal alteration, but RF and similar ensemble methods could be used to predict ore grades and the distribution of mineralisation. Other potential uses of

such multiparameter databases and artificial intelligence in exploration include: the prediction of lithology (pseudo-logs) along the boreholes, the generation of predictive maps of metals, resource estimation, and sample classification (e.g., Rodriguez-Galiano et al., 2015; Bérubé et al., 2018; Caté et al., 2018; Chen et al., 2018).

6. Conclusions

This study was made on a known VMS deposit, to validate the chosen method and then allow it to be used for exploration in Matagami or in other contexts. Our aim was to utilize machine learning to estimate an important chemical element, Na, which was missing from a multi-sensor core logging database. We first tested a Random Forest algorithm on our industrial partner's traditional lithogeochemistry data and obtained excellent results. We then applied the algorithm to the multiparameter data and obtained a usable estimate of Na at high spatial resolution. Random Forest is a very good tool when using large amounts of data and variables, whether relevant, or less relevant. However, it is important to have a sufficient number of relevant variables, in absolute terms and relative to the total number of variables. The RF method uses all variables, but automatically prioritizes the most useful, depending on the prediction requested. This method also allows visualization of the weight that each variable represents in the estimation. Future studies should compare RF with other methods.

Acknowledgements

This study summarizes the first author's MSc project at *Institut national de la recherche scientifique*. Funding came from a *Fonds de recherche du Québec - Nature et technologies* (FRQNT) grant to PSR and EG. Our industrial partner Glencore is acknowledged for site access, logistical support, and use of lithogeochemical data. Alexandre Bourke supervised the acquisition of multi-sensor data at Matagami and compiled the initial database. Field assistants were Catherine Frigon and Andréa Allard. We thank two anonymous journal reviewers for constructive comments on the manuscript.

References

Adair, R., 2009. Technical report on the resource calculation for the Bracemac-McLeod discoveries, Matagami project, Québec: National Instrument 43-101 report prepared on behalf of Donner Metals Ltd. (Vancouver, British Columbia), 132 p.

Bérubé, C. L., Olivo, G. R., Chouteau, M., Perrouty, S., Shamsipour, P., Enkin, R. J., Morris, W. A., Feltrin, L. & Thiémonge, R., 2018. Predicting rock type and detecting hydrothermal alteration using machine learning and petrophysical properties of the Canadian Malartic ore and host rocks, Pontiac Subprovince, Québec, Canada: Ore Geology Reviews, v. 96, p. 130-145.

Bourke, A. & Ross, P.-S., 2016. Portable X-ray fluorescence measurements on exploration drill-cores: comparing performance on unprepared cores and powders for 'whole-rock' analysis: Geochemistry: Exploration, Environment, Analysis, v. 16, p. 147-157.

Breiman, L., 2001. Random forests: Machine Learning, v. 45, p. 5-32.

Carranza, E. J. M. & Laborte, A. G., 2015. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines): Computers & Geosciences, v. 74, p. 60-70.

Caté, A., Perozzi, L., Gloaguen, E., Blouin, M., 2017. Machine learning as a tool for geologists. The Leading Edge, v. 36, p. 215-219.

Caté, A., Schetselaar, E., Mercier-Langevin, P. & Ross, P.-S., 2018. Classification of lithostratigraphic and alteration units from drillhole lithogeochemical data using machine learning: A case study from the Lalor volcanogenic massive sulphide deposit, Snow Lake, Manitoba, Canada. Journal of Geochemical Exploration, v. 188, p. 216-228.

Chen, S., Hattori, K. & Grunsky, E. C., 2018, Identification of sandstones above blind uranium deposits using multivariate statistical assessment of compositional data, Athabasca Basin, Canada: Journal of Geochemical Exploration, v. 188, p. 229-239.

Debreil, J.-A., Ross, P.-S. & Mercier-Langevin, P., 2018. The Matagami district, Abitibi Greenstone Belt, Canada: volcanic controls on Archean volcanogenic massive sulfide deposits associated with voluminous felsic volcanism: Economic Geology, v. 113, p. 891-910.

Fischer, A., 2014. Deux méthodes d'apprentissage non supervisé: synthèse sur la méthode des centres mobiles et présentation des courbes principales: Journal de la Société Française de Statistique, v. 155, no. 2, p. 2-35.

Franklin, J. M., Gibson, H. L., Jonasson, I. R. & Galley, A. G., 2005. Volcanogenic massive sulfide deposits. In J. W. Hedenquist, J. F. H. Thompson, R. J. Goldfarb & J. P. Richards (Eds.), Economic Geology One Hundredth Anniversary Volume, Society of Economic Geologists, p. 523-560.

Fresia, B., 2013. Analyses multivariées de données de forage de la région de Matagami: MSc thesis, Institut national de la recherche scientifique, Québec, Canada, 152 p.

Fresia, B., Ross, P.-S., Gloaguen, E. & Bourke, A., 2017. Lithological discrimination based on statistical analysis of multi-sensor drill core logging data in the Matagami VMS district, Quebec, Canada: Ore Geology Reviews, v. 80, p. 552–563.

Genna, D., Gaboury, D. & Roy, G., 2014. The Key Tuffite, Matagami Camp, Abitibi Greenstone Belt, Canada: petrogenesis and implications for VMS formation and exploration: Mineralium Deposita, v. 49, p. 489-512.

Gifkins, C., Herrmann, W. & Large, R. R., 2005. Altered volcanic rocks: A guide to description and interpretation, Centre for Ore Deposit Research, University of Tasmania, 275 p.

Heutte, L., Bernard, S., Adam, S. & Oliveira, É., 2008. De la sélection d'arbres de décision dans les forêts aléatoires, *in* Proceedings, Colloque International Francophone sur l'Écrit et le Document, Groupe de Recherche en Communication Écrite, p. 163-168.

Jácomo, M. H., Brod, T. C. J., Pires, A. C. B., Brod, J. A., Palmieri, M. & de Melo, S. S. V., 2015. Magnetic susceptibility and gamma-ray spectrometry on drill core: lithotype characterization and 3D ore modelling of the Morro Do Padre niobium deposit, Goiás, Brazil: Revista Brasileira de Geofísica, v. 33, p. 15.

Large, R. R., Gemmell, J. B., Paulick, H. & Huston, D. L., 2001. The alteration box plot: A simple approach to understanding the relationship between alteration

mineralogy and lithogeochemistry associated with volcanic-hosted massive sulfide deposits: Economic geology, v. 96, p. 957-971.

Martín-Fernández, J. A., Barceló-Vidal, C. & Pawlowsky-Glahn, V., 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation: Mathematical Geology, v. 35, p. 253-278.

Mercier-Langevin, P., Gibson, H. L., Hannington, M. D., Goutier, J., Monecke, T., Dubé, B. & Houlé, M. G., 2014. A special issue on Archean magmatism, volcanism, and ore deposits: part 2. Volcanogenic massive sulfide deposits preface. Economic Geology, v. 109, p. 1-9.

Mohamadally, H. & Fomani, B., 2006. SVM : Machines à vecteurs de support ou séparateurs à vastes marges : Rapport technique, Versailles St Quentin, France, 64 p.

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

Piché, M., 1991, Synthèse géologique et métallogenique du camp minier de Matagami, Québec. Unpublished PhD thesis, Université du Québec à Chicoutimi, Canada, 249 p.

Raschka, S, & Mirjalili, V., 2017. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition. Packt Publishing, 622 p.

Rakotomalala, R., 2005. Arbres de décision: Revue Modulad, v. 33, p. 163-187.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. & Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines: Ore Geology Reviews, v. 71, p. 804-818.

Ross, P.-S. & Bourke, A., 2017. High-resolution gamma ray attenuation density measurements on mining exploration drill cores, including cut cores: Journal of Applied Geophysics, v. 136, p. 262-268.

Ross, P.-S., Bourke, A. & Fresia, B., 2013. A multi-sensor logger for rock cores: Methodology and preliminary results from the Matagami mining camp, Canada: Ore Geology Reviews, v. 53, p. 93-111.

Ross, P.-S., Bourke, A. & Fresia, B., 2014a. Improving lithological discrimination in exploration drill-cores using portable X-ray fluorescence measurements: (1) testing three Olympus Innov-X analysers on unprepared cores: Geochemistry: Exploration, Environment, Analysis, v. 14, p. 171-185.

Ross, P.-S., Bourke, A. & Fresia, B., 2014b. Improving lithological discrimination in exploration drill-cores using portable X-ray fluorescence measurements: (2) applications to the Zn-Cu Matagami mining camp, Canada: Geochemistry: Exploration, Environment, Analysis, v. 14, p. 187-196.

Ross, P.-S., Bourke, A., Mercier-Langevin, P., Lépine, S., Leclerc, F. & Boulerice, A., 2016a. High-resolution physical properties, geochemistry and alteration mineralogy for the host rocks of the Archean Lemoine auriferous VMS deposit, Canada. Econ. Geol., v. 111, p. 561-1574.

Ross, P.-S., Schnitzler, N. & Bourke, A., 2016b. Analyse multiparamétrique à haute résolution de carottes de forage dans la région de Matagami 2014-2015, résultats préliminaires: Ministère de l'Énergie et des Ressources naturelles du Québec, report MB 2016-17, 60 p.

Roy, G. & Allard, M., 2006. Matagami, une approche ciblée sur de nouveaux concepts: Québec Exploration 2006, Ministère des Ressources naturelles et de la Faune (Québec), document DV 2006, p. 13.

Schnitzler, N., 2017. Apprentissage automatique de données mutiparamétriques au gisement de sulfures massifs volcanogènes Bracemac-McLeod, district minier de Matagami, Québec. MSc thesis, Institut national de la recherche scientifique, Québec, 115 p.

Sharpe, J., 1968. Geology and sulfide deposits of the Matagami area, Abitibi-East County. Ministère des Richesses Naturelles du Québec, report RG-137(A), 130 p.

Wang, R., Cudahy, T., Laukamp, C., Walshe, J. L., Bath, A., Mei, Y., Young, C., Roache, T. J., Jenkins, A. J., Roberts, M., Barker, A. & Laird, J., 2017. White mica as a hyperspectral tool in exploration for the Sunrise Dam and Kanowna Belle gold deposits, Western Australia: Economic Geology, v. 112, p. 1153-1176.

Figures

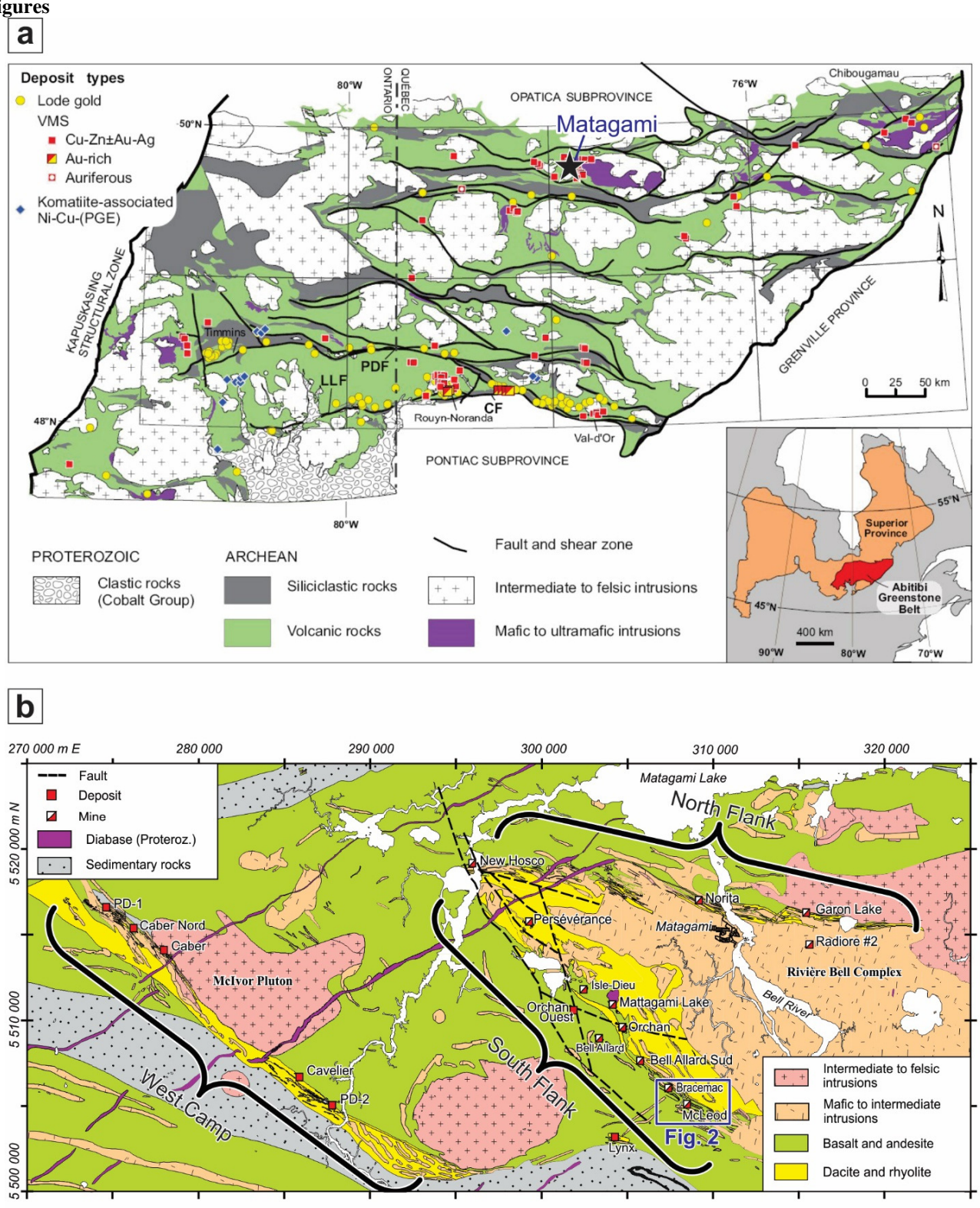

Figure 1: Geological context of the Matagami mining district. (a) Simplified geological map of the Abitibi Subprovince (after Mercier-Langevin et al., 2014), showing the position of the Matagami district (black star). Inset at lower right shows the Abitibi Subprovince within the Superior Province. (b) Simplified geological map of the Matagami area, after Roy and Allard (2006). UTM grid is NAD 83, zone 18.

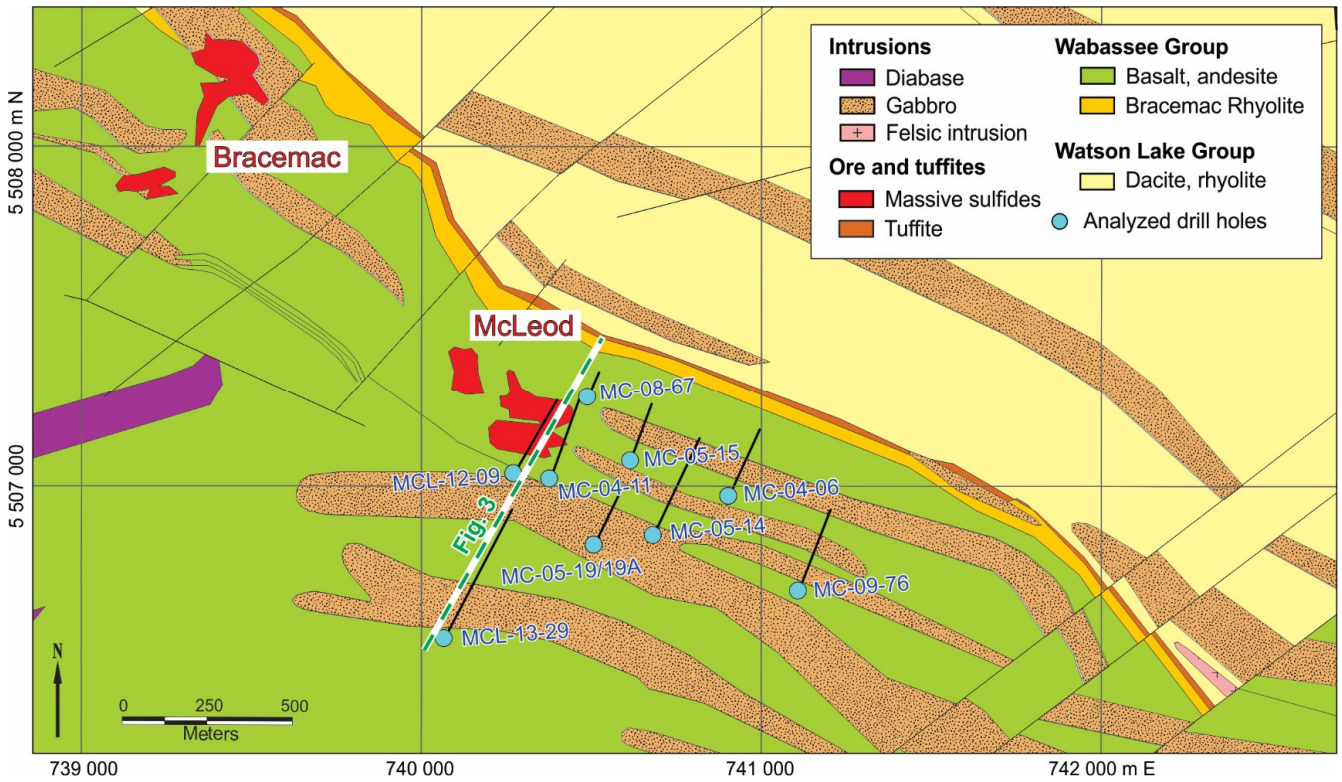

Figure 2: Geological map of the Bracemac-McLeod area, compiled par Glencore, showing the distribution of the nine analyzed drill holes used in this study. Note that the ore lenses are projected vertically to the surface, whereas the volcanic strata dip steeply to the SW. UTM grid is NAD 83, zone 17.

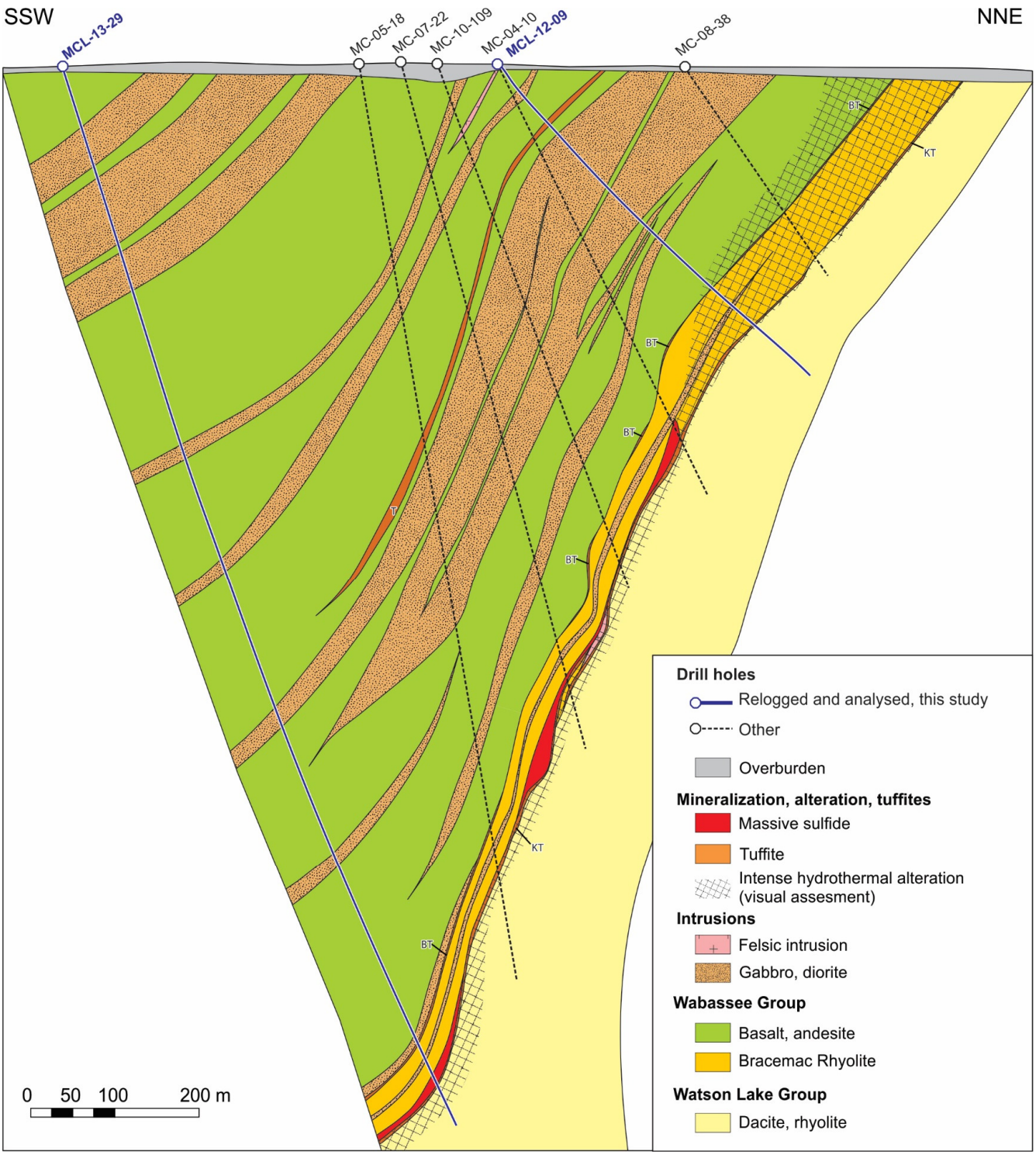

Figure 3: Vertical section passing through the McLeod deposit (see Fig. 2), showing the typical geological sequence of the area. The two drill holes used in this study are in blue. BT= Bracemac Tuffite, KT = Key Tuffite.

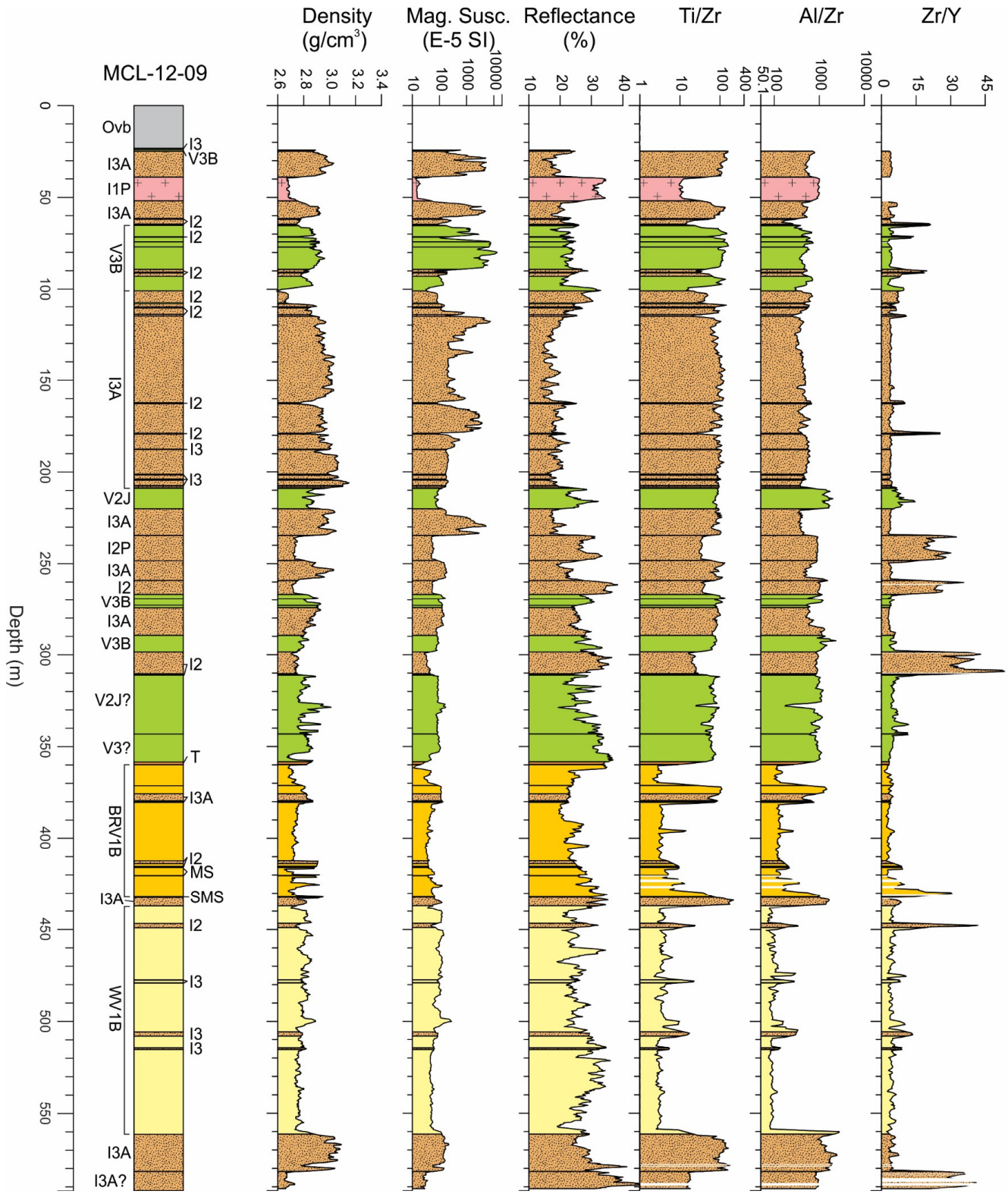

Figure 4: Physical properties and geochemical ratios as a function of depth for drillhole MCL-12-09. Blank areas: concentration is below the limit of detection. The geology log was produced by the first author. Lithological codes: BRV1B = Bracemac Rhyolite; I1P = felsic intrusion, porphyritic; I2, I2P = intermediate intrusion (P = porphyritic); I3, I3A = gabbro; MS = massive sulfides; Ovb = overburden; SMS = semi-massive sulfides; T = tuffite; V2J, V3, V3B = intermediate to mafic lavas, undifferentiated; WV1B = Watson Lake Rhyolite.

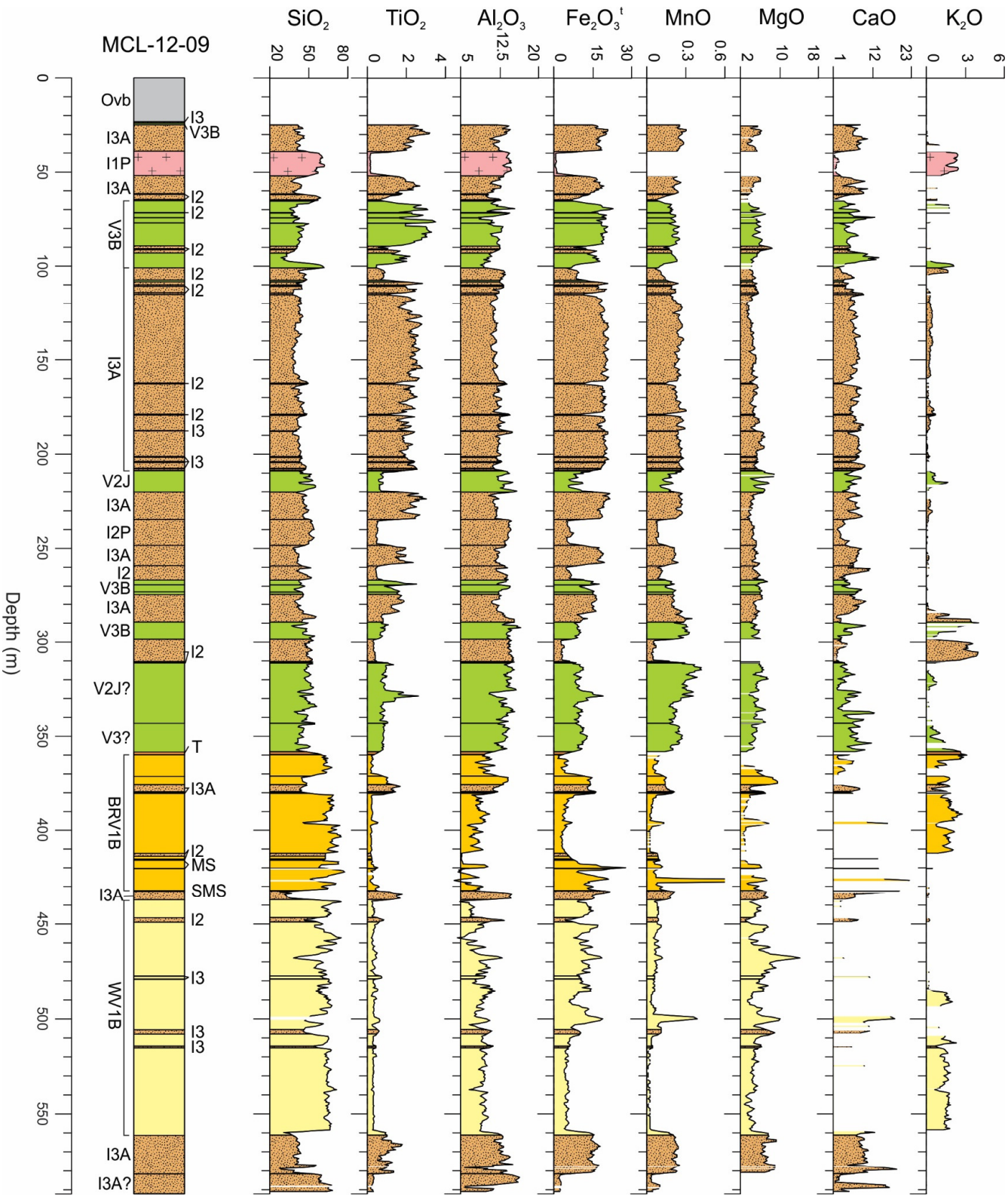

Figure 5: Corrected major oxides (%) as a function of depth for drillhole MCL-12-09. See figure 4 for other explanations.

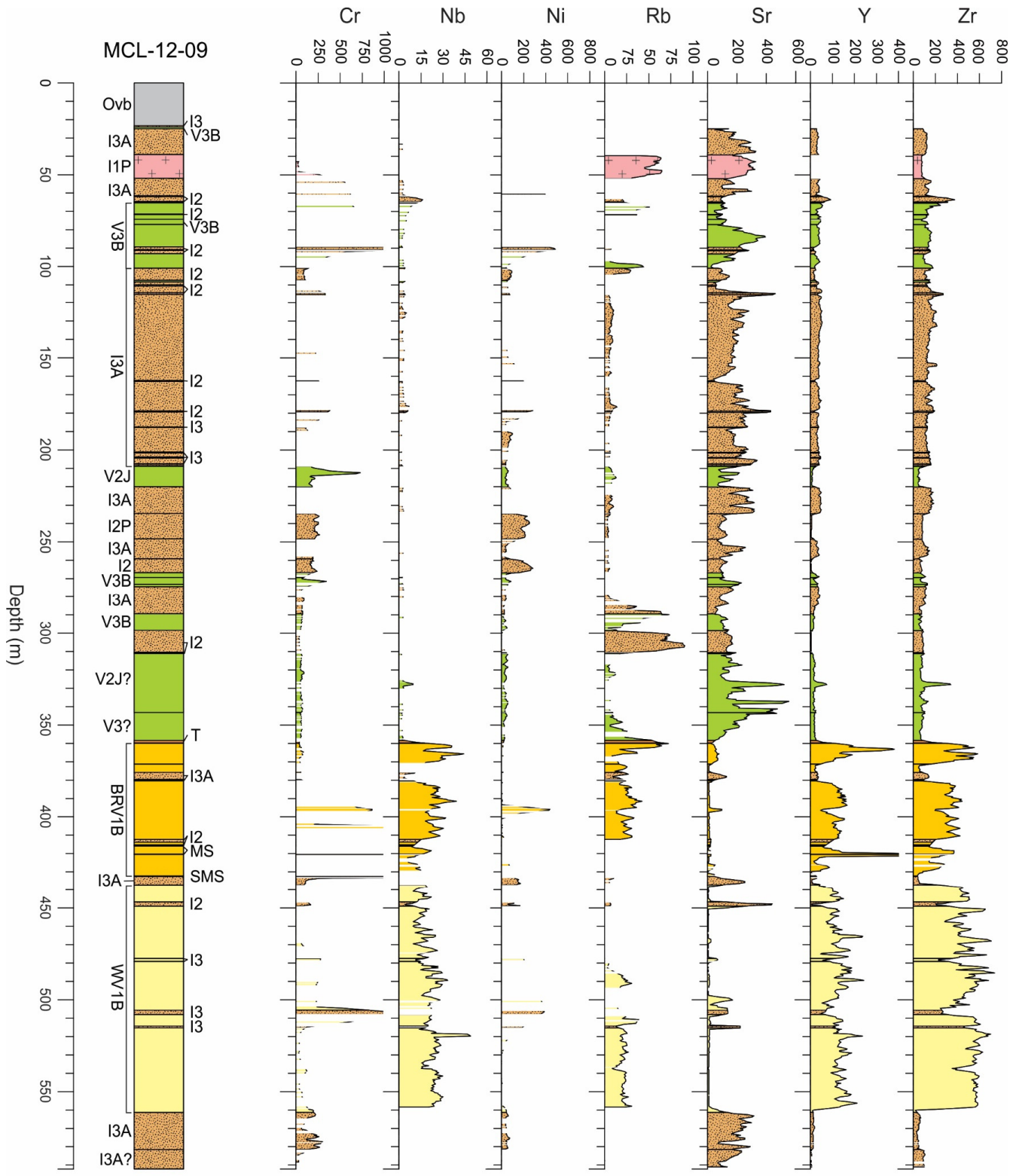

Figure 6: Corrected trace elements (ppm) as a function of depth for drillhole MCL-12-09. See figure 4 for other explanations.

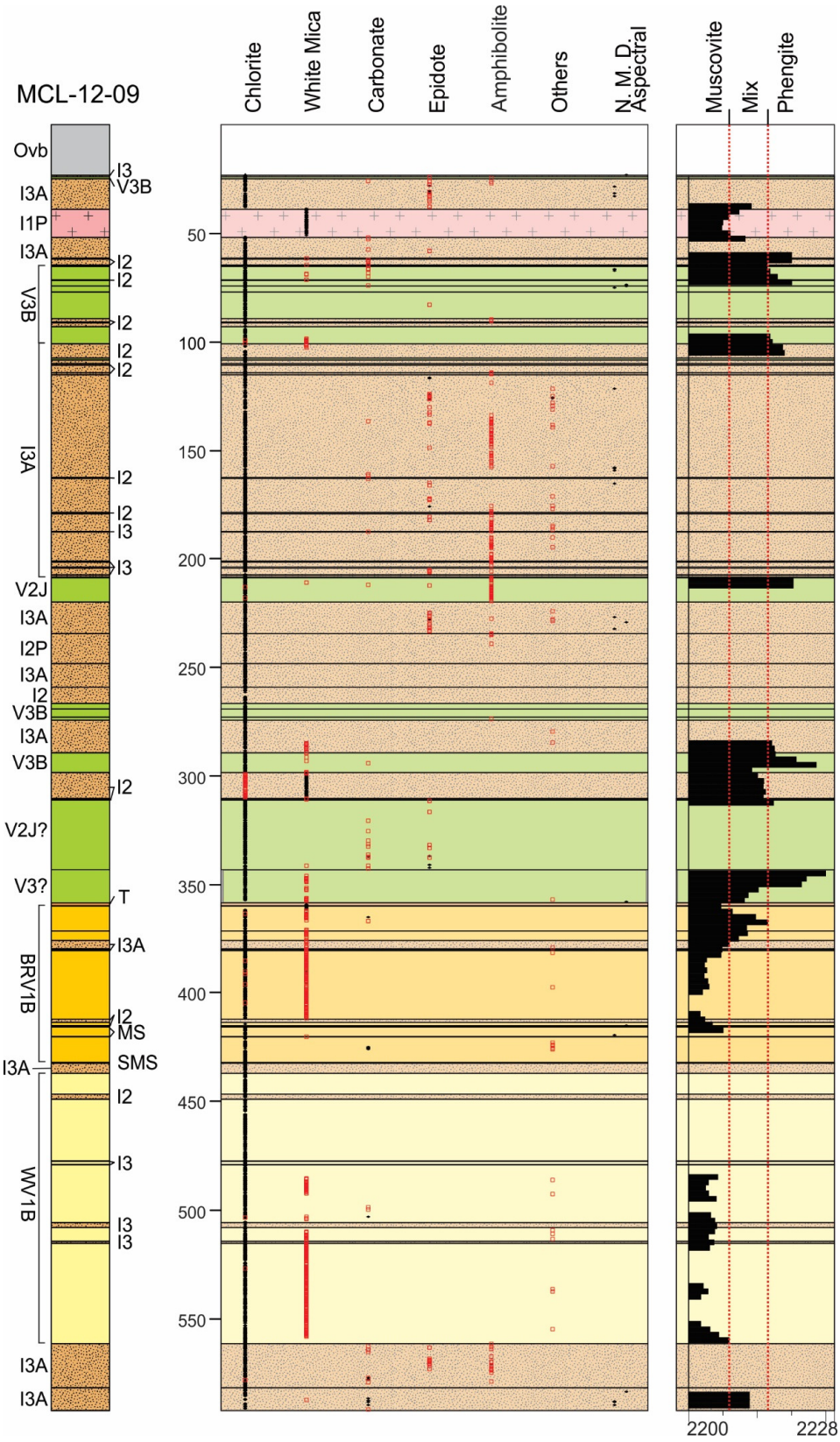

Figure 7: Infrared mineralogy as a function of depth for drillhole MCL-12-09. A small black diamond is the main mineral group (“Group 1”), and a small red square is the second mineral group (“Group 2”) that explains each infrared spectrum. "N.M.D." = no mineral detected. Right part of figure shows the composition of white mica, based on the wavelength of the AlOH peak, in nanometers.

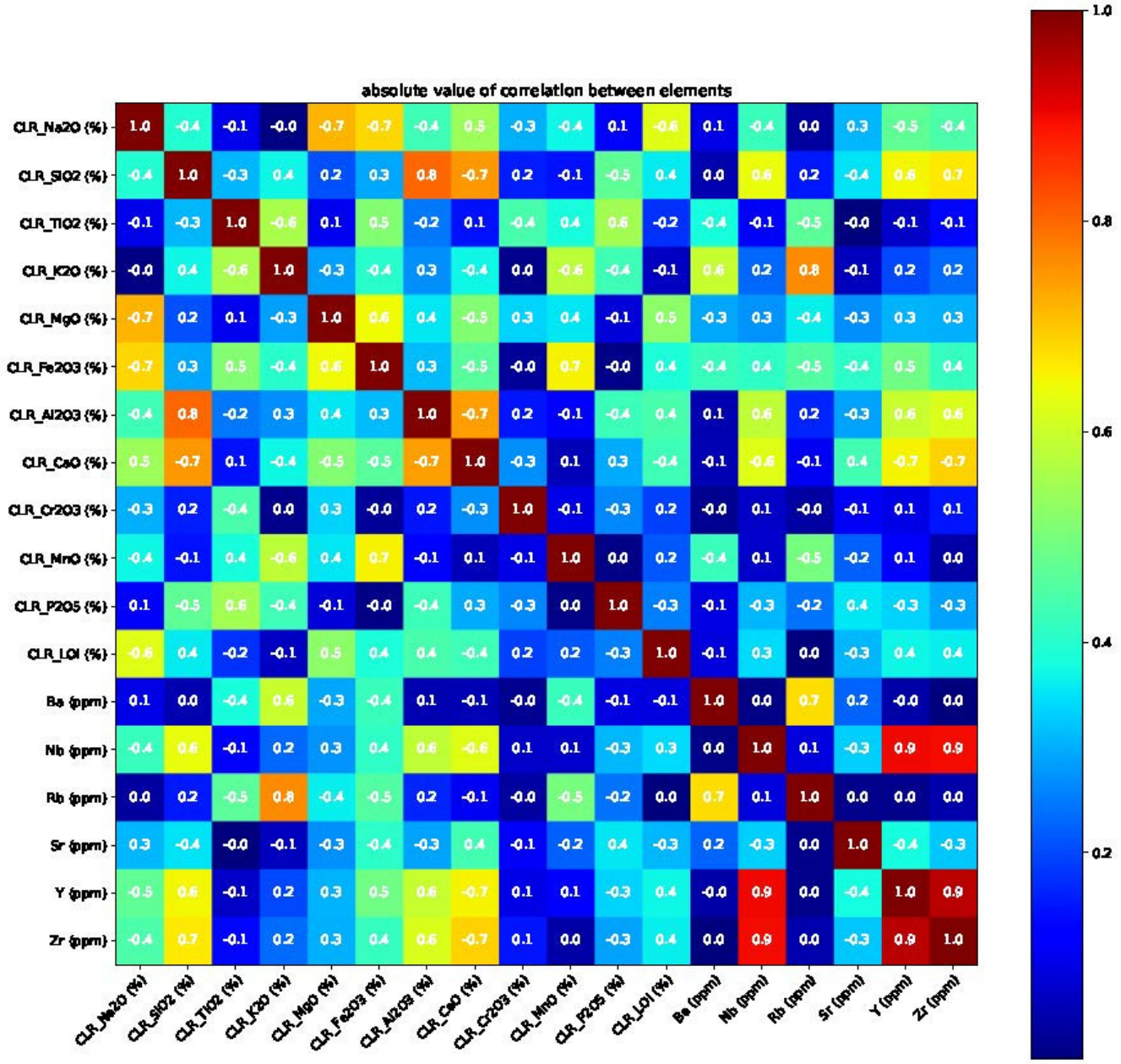

Figure 8: Correlation matrix for the Glencore geochemical data in the Bracemac- McLeod sector. The centered log ratio data is used for the major oxides.

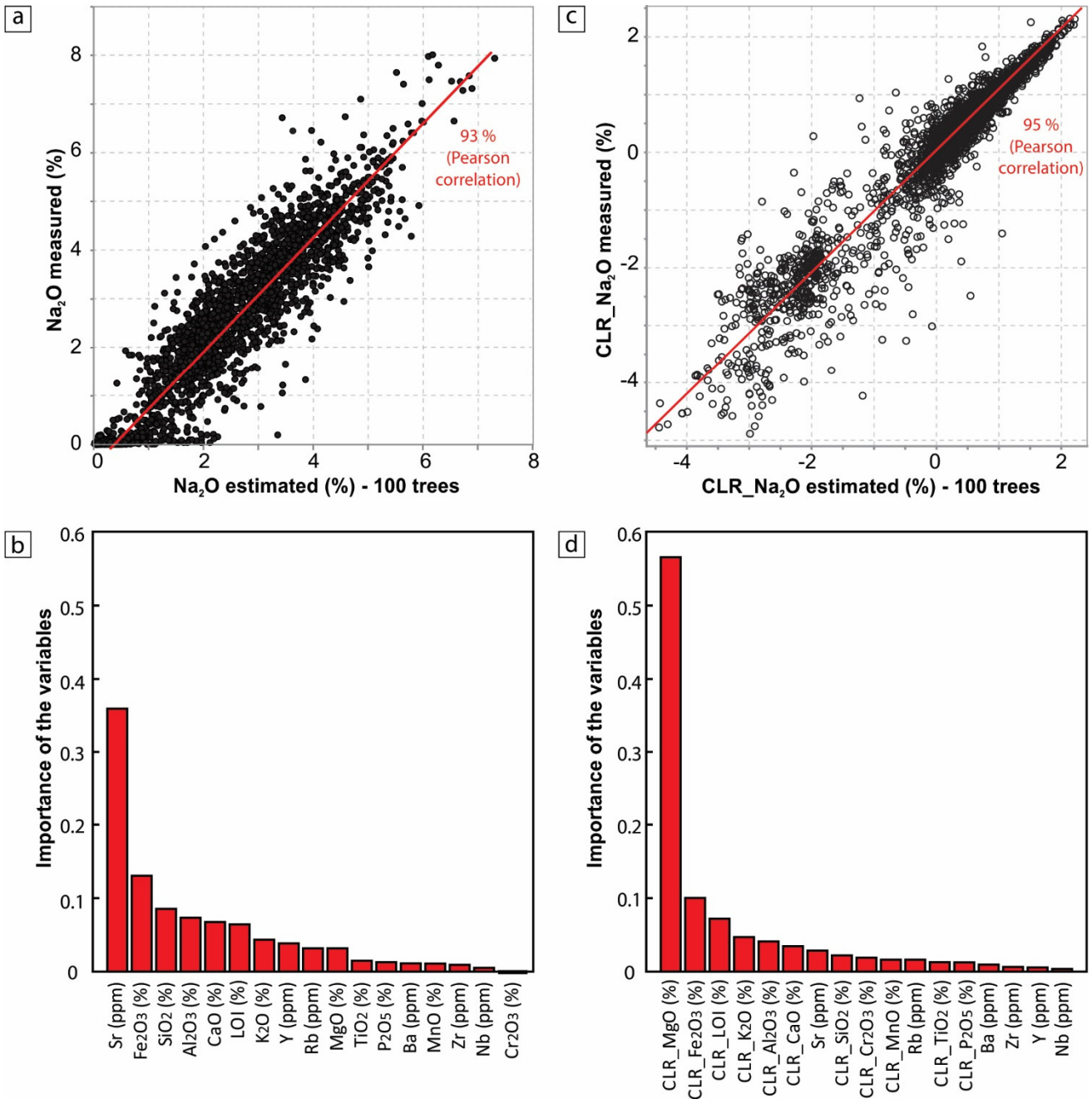

Figure 9: Step 1, estimating Na with the Glencore geochemical data in the Bracemac- McLeod sector, 314 drill holes, $n_{variable}$ = 18, result of the test subset (n_{data} = 2763). (a) Estimated versus measured Na using untransformed data. (b) Importance of variables for the estimation shown in (a). (c) Estimated versus measured Na using centered log ratio data. (d) Importance of variables for the estimation shown in (c).

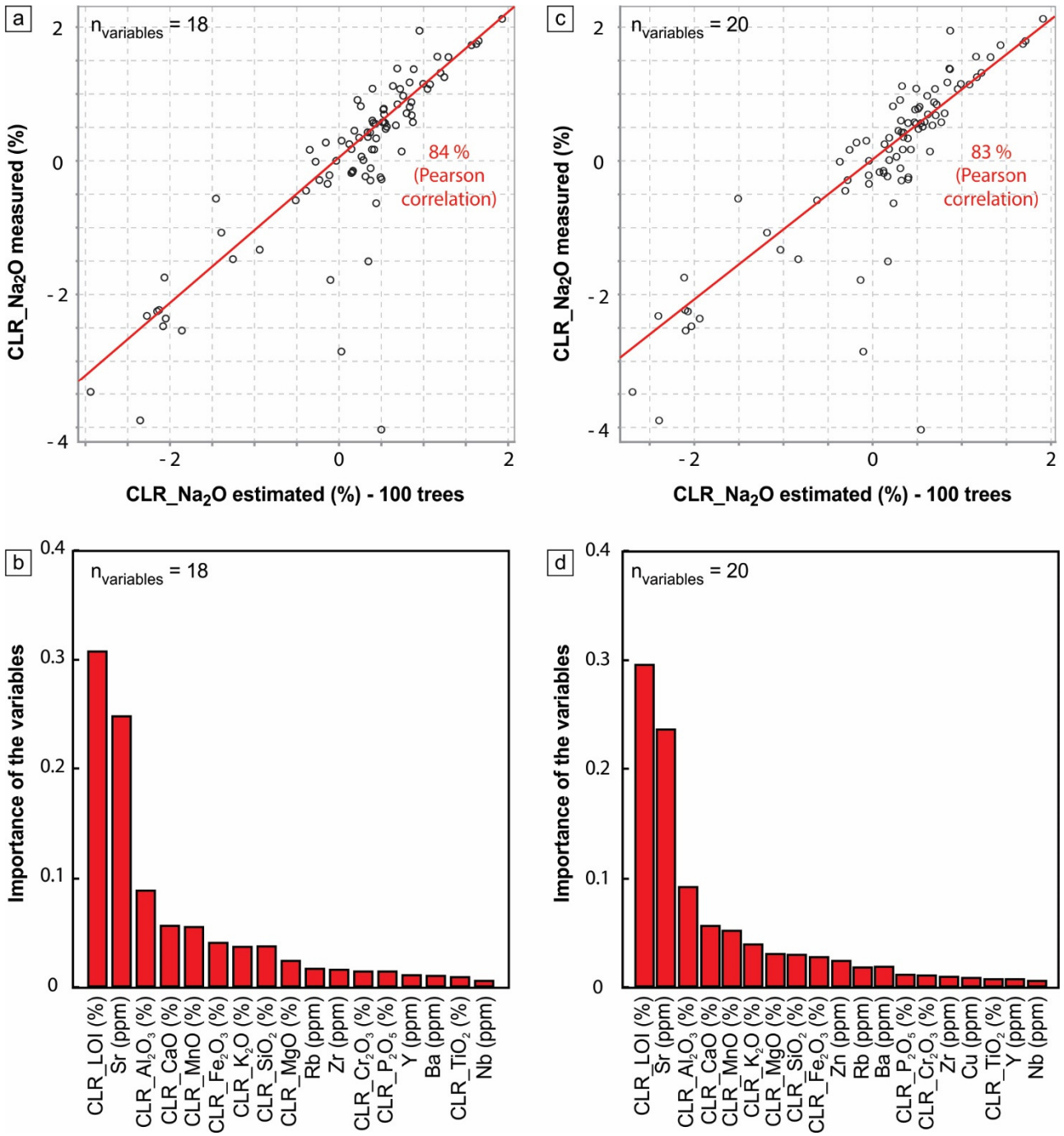

Figure 10: Step 2, estimating Na with the Glencore geochemical data, nine holes only, centered log ratio transformation, results of the test subset (n_{data} = 87). (a) Estimated versus measured Na, $n_{variable}$ = 18. (b) Importance of variables for the estimation shown in (a). (c) Estimated versus measured Na, $n_{variable}$ = 20. (d) Importance of variables for the estimation shown in (c).

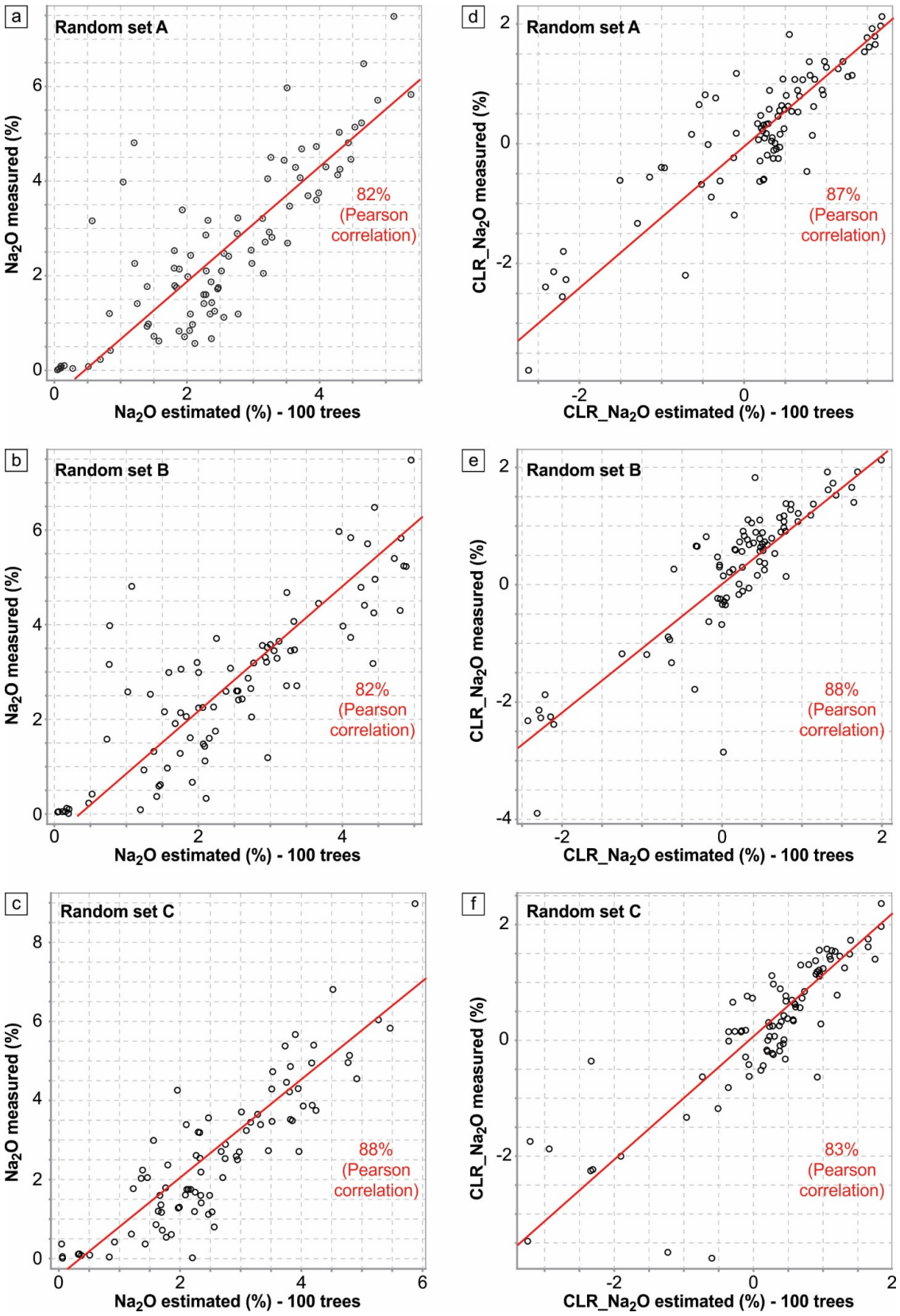

Figure 11: Step 3, testing the influence of the training set, with the Glencore geochemical data for the nine drill holes, $n_{variable}$ = 20. Results (estimated versus measured Na) are shown for three random sets (called A, B and C). (a)-(b)-(c) Untransformed data. (d)-(e)-(f) Centered log ratio data.

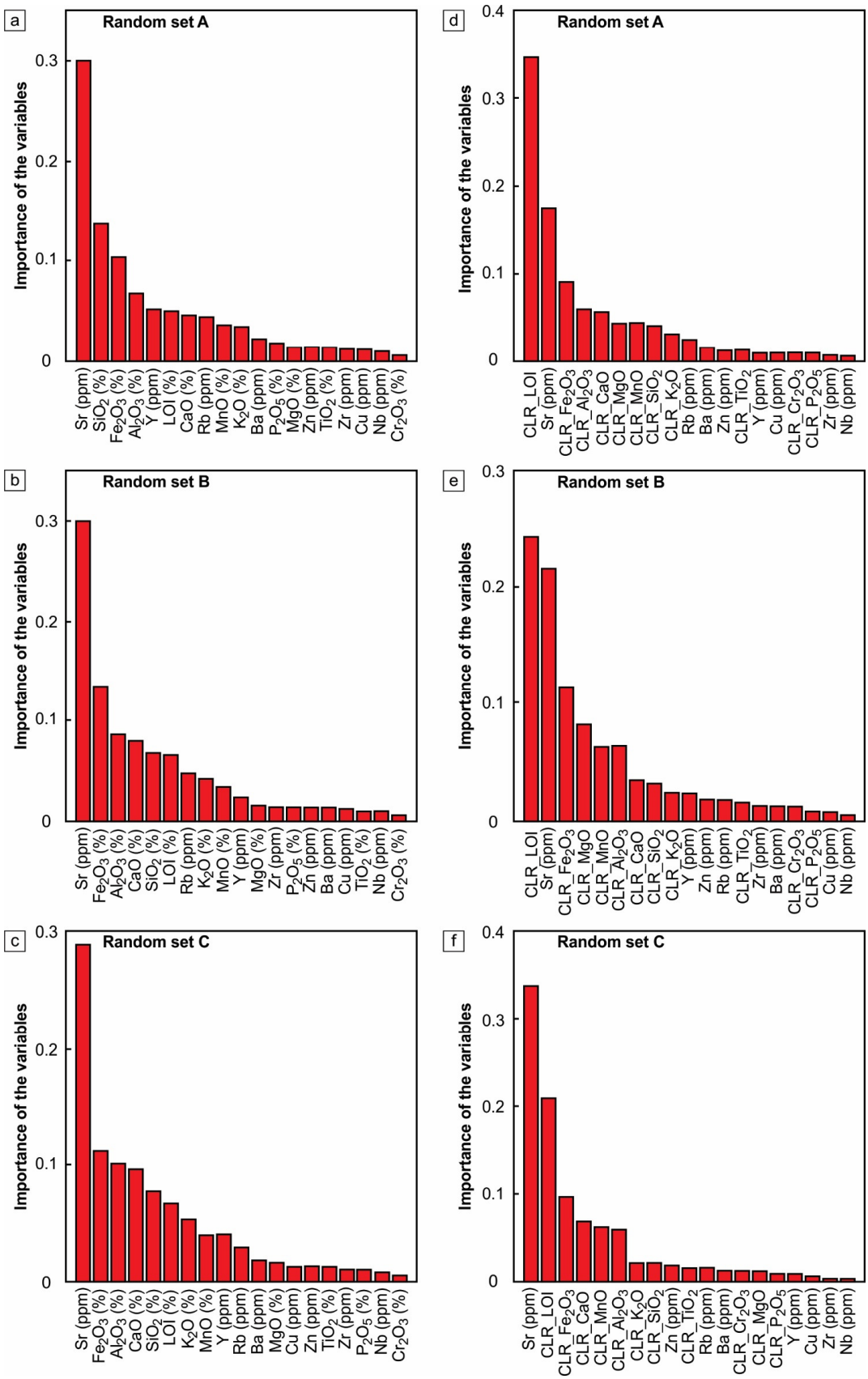

Figure 12: Step 3, testing the influence of the training set with three random subsets, continued: importance of variables for the estimations shown in figure 11. (a)-(b)-(c) Untransformed data. (d)-(e)-(f) Centered log ratio data.

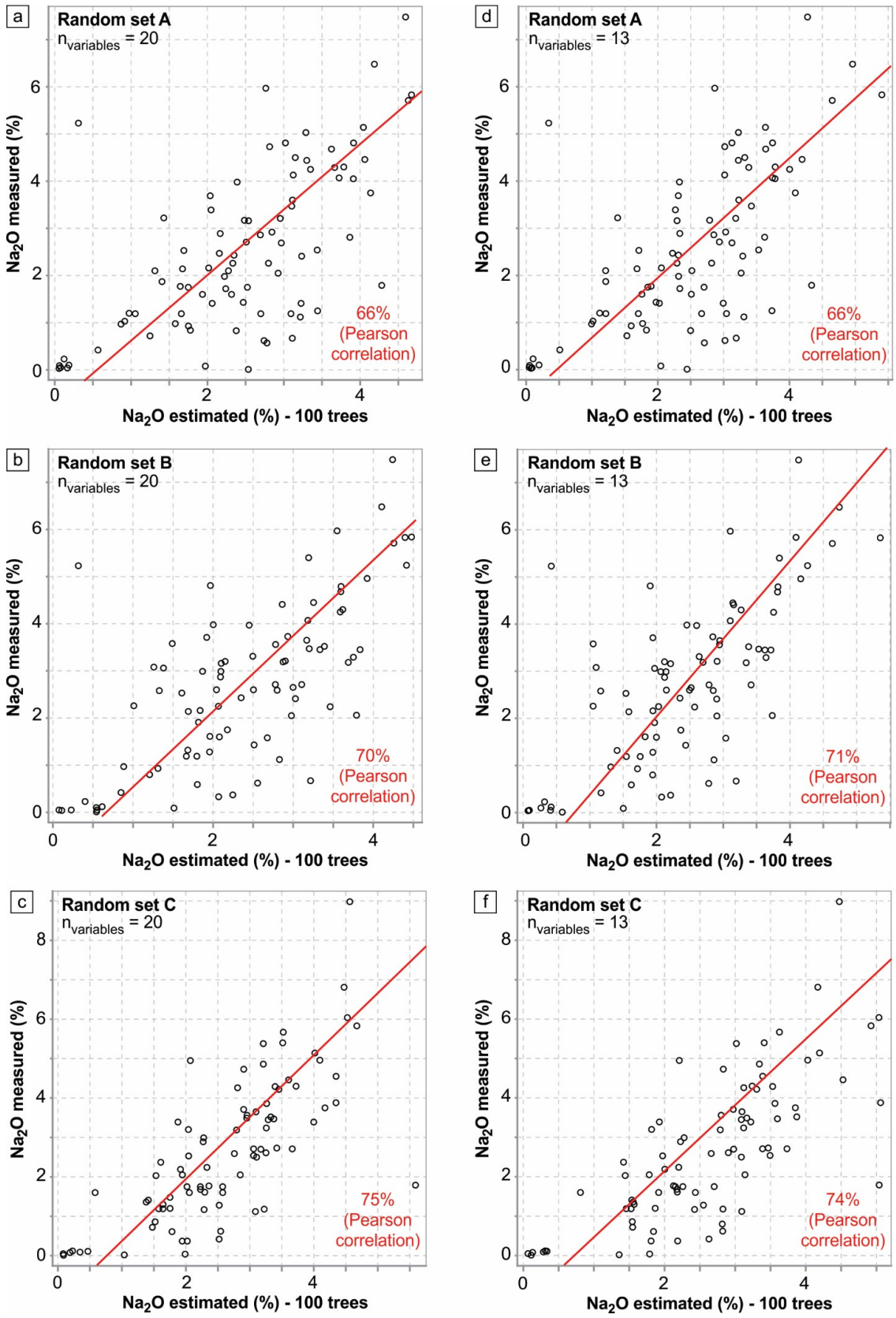

Figure 13: Step 4, estimating Na with the untransformed LAMROC data, with the measured Na borrowed from adjacent Glencore geochemical analyses, for three random sets (called A, B and C). (a) (b)-(c) $n_{variable}$ = 20. (d)-(e)-(f) $n_{variable}$ = 13 (suppression of elements with more than 10% of data below the detection limit).

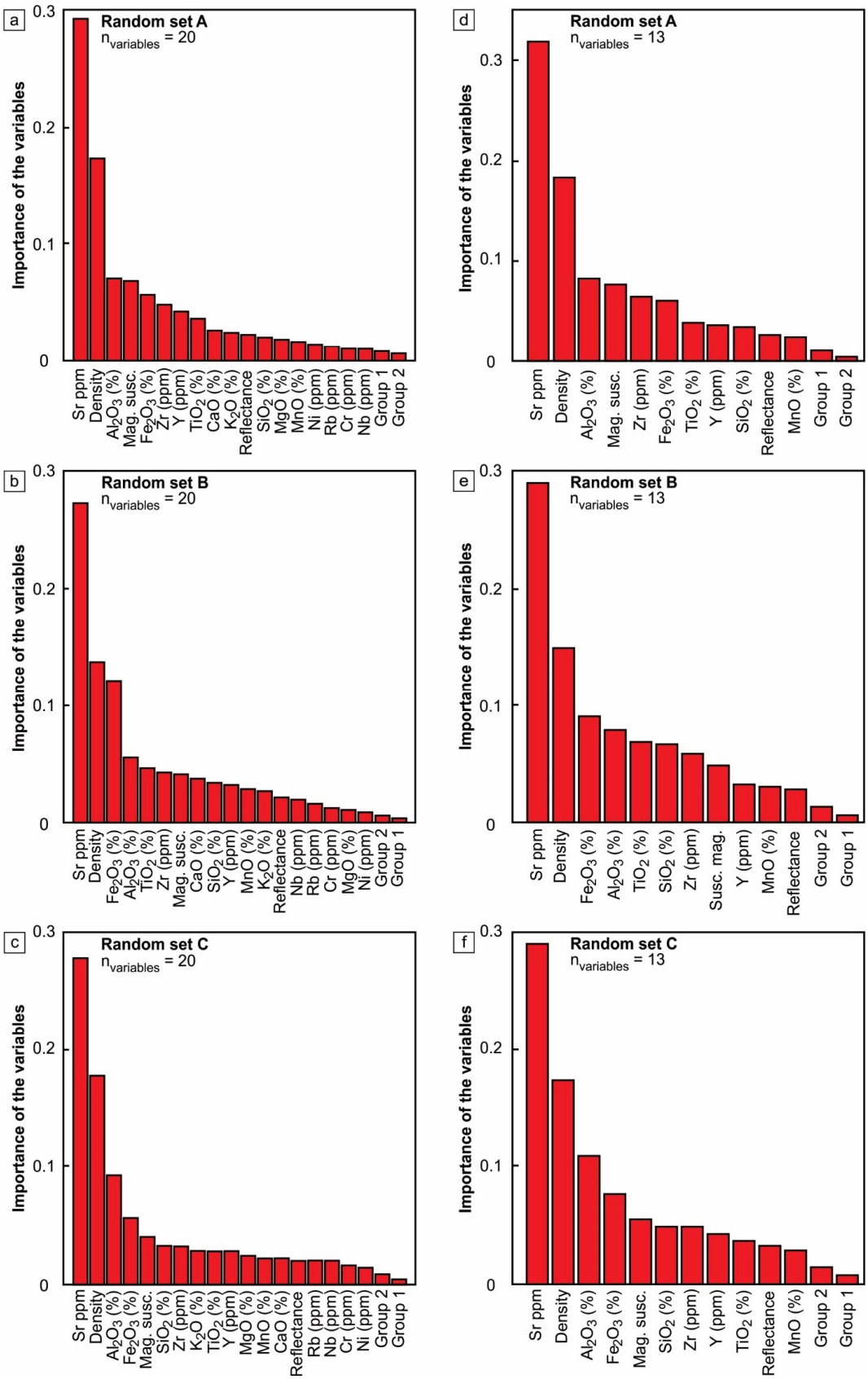

Figure 14: Step 4, estimating Na with the untransformed LAMROC data, continued: importance of variables for the estimations shown in figure 13. (a) (b)-(c) $n_{variable}$ = 20. (d)-(e)-(f) $n_{variable}$ = 13.

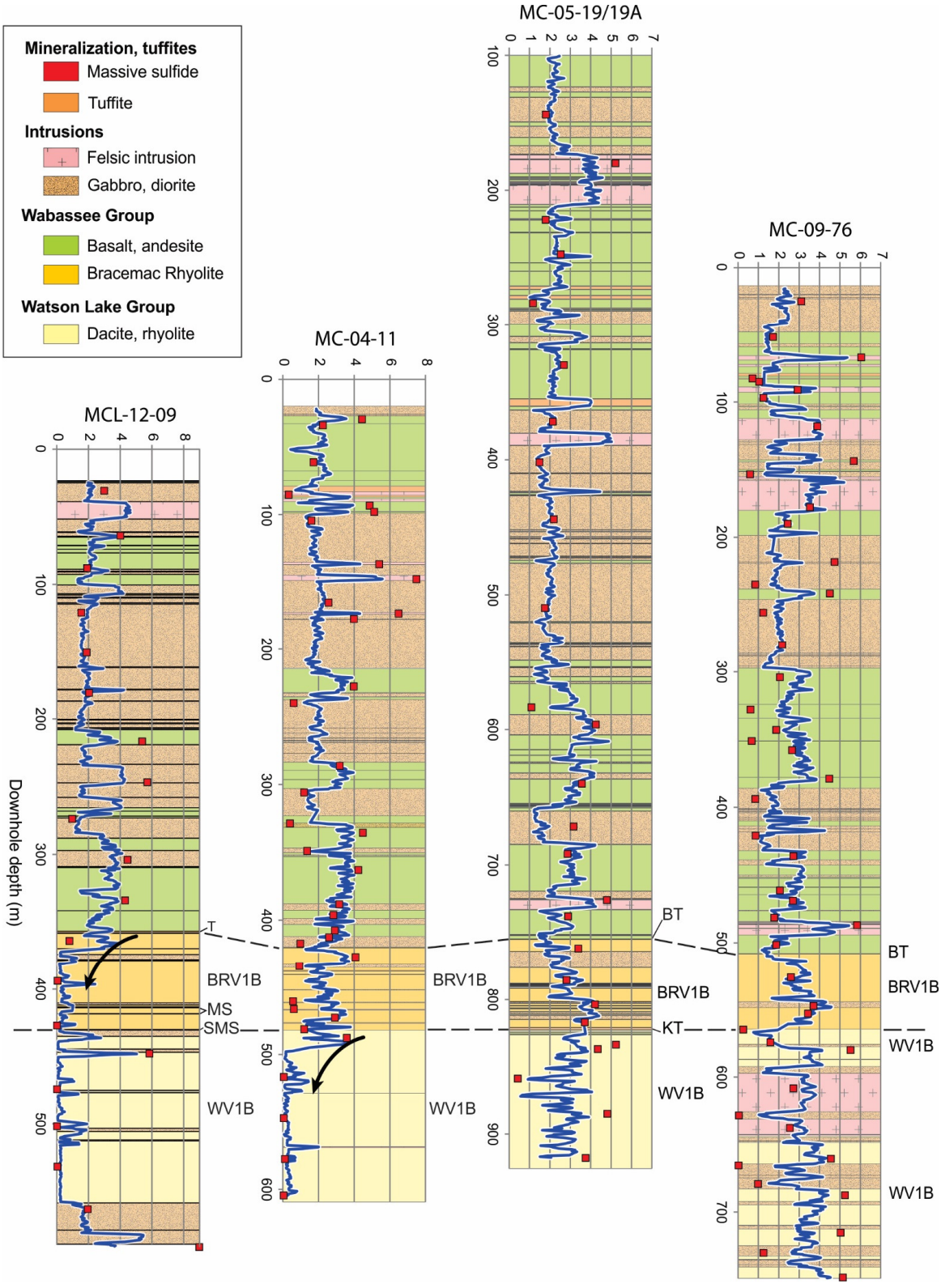

Figure 15: Profiles of sodium versus depth for four drill holes (MCL-12-09, MC-04-11, MC-05-19/19A and MC-09-76) arranged from proximal to distal relative to the McLeod VMS deposit (see Fig. 2 for locations). Blue line shows estimated sodium based on LAMROC data. Red squares are Glencore geochemical data. BRV1B = Bracemac Rhyolite; BT = Bracemac Tuffite; KT = Key Tuffite; MS = massive sulfide; SMS = semi-massive sulfide; T = tuffite; WV1B = Watson Lake Rhyolite. The profiles are vertically arranged so that the KT horizon is always at the same level. Downhole depth does *not* represent true thickness.

Tables

Table 1: Summary of the steps performed to estimate sodium.

Raw data	Step	Goal	Number of drillholes	Number of data	Number of variables	Data type (unprocessed or centered log-ratio)	Results/Pearson correlation	Most influential variables
Glencore	Step 1	Test algorithm on large database of conventional geochemistry	314	8287	18	Unprocessed data	0.93	Sr (35%); Fe_2O_3 (13%)
			314	8287	18	CLR	0.95	MgO (57%); Fe_2O_3 (10%)
	Step 2	Test algorithm with less data	9	260	18	CLR	0.84	LOI (31%); Sr (25%)
		Test algorithm with less data, but adding two variables	9	260	20		0.83	LOI (30%); Sr (24%)
	Step 3	Vary training and test sets (A, B and C)	9	260	20	Unprocessed data	0.82-0.88	Sr (30%); SiO_2 (10%) (A) Sr (30%); Fe_2O_3 (10%) (B and C)
						CLR	0.82-0.88	LOI (25-35%); Sr (20%) (A and B) Sr (35%); LOI (20%) (C)
LAMROC	Step 4	Variation of variables and training and test sets (A, B and C)	9	260	20	Unprocessed data	0.66-0.75	Sr (30%); density (12-18%) (A, B and C)
			9	260	20	CLR	0.66-0.74	Sr (30-35%); density (15-18%) (A, B and C)
			9	260	13	Unprocessed data	0.69-0.75	Sr (40%); density (10%) (A and C) Sr (40%); Al_2O_3 (15%) (B)
	Step 5	Estimate Na based on full multivariate dataset	9	9675	20	Unprocessed data	-	-