AN EFFICIENT WORKFLOW FOR PREDICTING VARIOUS LOGGING VARIABLES USING SIMPLE MACHINE-LEARNING PROGRAMS

M. SERDOUN
GeoRessources, CNRS, Université de Lorraine, Labcom CREGU
Vandoeuvre-lès-Nancy, France
mehdi.serdoun@univ-lorraine.fr

F. SUR
LORIA, Université de Lorraine, Inria, CNRS,
Vandœuvre-lès-Nancy, France

E. WILLIARD
ORANO
Châtillon, France

P. LEDRU
GeoRessources, CNRS
Université de Lorraine, Labcom CREGU
France

T. OBIN
GeoRessources, CNRS
Université de Lorraine, Labcom CREGU
France

G. MILESI
GeoRessources, CNRS, Université de Lorraine
Labcom CREGU
France

A. DONEY
ORANO Canada
Saskatoon, Canada

LE BEUX
ORANO Châtillon
France

J. MERCADIER
GeoRessources, CNRS, Université de Lorraine
Labcom CREGU
France

Abstract

The paper presents simple tools for the prediction of logging variables in uranium exploration using various instrumental data. These tools include i) the prediction of potassic alteration using routine IR-spectroscopy in the 350-2500nm range using Partial Least Squares Regression (PLS-R), ii) the prediction of alteration facies using Visible-NIR (350-1000nm) spectroscopy along with Partial Least Squares – Discriminant Analysis (PLS-DA) and iii) lithostratigraphic units classification using geochemical assays along with a Random Forest Classifier. These tools are associated with an open online repository describing a standard Machine-Learning pipeline for drilling data.

1. WHY MACHINE-LEARNING FOR DRILLHOLE DATA?

As easy-to-explore mineral deposits tend to progressively be depleted in most greenfield exploration districts, exploitable deposits tend to be located under cover at increasing depths within the earth's crust. Exploration for uranium deposits in the Athabasca Basin (Saskatchewan, Canada) is no exception to this general trend as exemplified by the recent discoveries of deeply-seated deposits of Arrow (discovered in 2011, 1000m), Phoenix (2009, 400m), Fox Lake (2010, 700m) or Centennial (2005, 800m). As a direct consequence, exploration of uranium in the Athabasca Basin like other commodities is becoming more and more costly over time without significant improvements in the rate of new discovery. All other things being equal, it was estimated in 2016 that the unit discovery cost - i.e., amount of money spent in exploration per quantity of U discovered - had to be divided by two to four for the industry to be sustainable over the coming decades [1].

In this paradoxical context and with the rapid and recent improvement of multiple analytical methods integrated into exploration procedures, recent years have seen an explosion in the availability of data produced from drilling for most exploration prospects, notably with the generalization of handheld real-time acquisition methods (e.g. XRF, IR spectroscopy), new geophysical methods, and the increasing accuracy and number of elements measured in geochemical assays at labs (i.e., ICP-MS). Machine learning (ML)– a general term encompassing a set of algorithmic methods aimed at retrieving statistical patterns and predicting unknown variables from multivariate datasets – provides an opportunity for valuing this increasing quantity of information, for helping to model more complex geological processes and for providing geologists with rapid feedback on the hidden relationships lying within exploration data.

Uranium mining companies can draw many direct benefits from the efficient use of ML related methods in the valorization of their drilling data, including:

- A decrease in exploration costs: some of the most valuable types of data to exploration geologists are either time-consuming to acquire, expensive or both, whilst other types of data are faster and cheaper to obtain. Most ML algorithms are nowadays free-to-use and widely available to the geoscientific community, and a good use of ML algorithms along with these fast and cheap methods for the prediction of more expensive ones thus represents an important asset in exploration strategies.
- Decrease human bias and error in exploration: manual reporting of variables of interest for exploration is highly time-consuming and potentially biased due to variations in the interpretations made by geologists. Various Machine-Learning and chemometric approaches, including multivariate calibration methods (like PLS-DA and PLS-R) can help to produce automated logs integrating information from multiple datasets.
- Historical prospects relogging and mineral potential reassessment: past experience has shown that numerous economic deposits are probably located on prospects already thoroughly explored and incorrectly thought to be barren. When the logging and acquisition of instrumental data have been missing, inconsistent or poorly recorded in older prospects compared to current standards, ML represents a valuable tool in the reassessment of the mineral potential of these areas through completion of missing data without the need to spend time and money re-exploring them.

The objective of this contribution is not to present purely optimal (although care is given about optimization problems to some extent) ML programs that can only be only understood and used by people well-versed in algorithms design, but rather to give near-optimal algorithms that require little computing time, power, and programming knowledge to be implemented and can be readily used by geologists during field work or for reassessment of the potential of historical prospects.

Three examples from real-world drillhole data are given, allowing the prediction of i) alteration facies using a chemometric method for classification (Partial Least Squares – Discriminant Analysis), ii) potassic alteration/K_2O content using a chemometric method for regression (Partial Least Squares - Regression), and iii) stratigraphic units using lithogeochemical assays. All examples were designed using exploration data that were not originally collected

with calibration problems in mind and problems related to reconciliation of unevenly sampled data were questioned. An open github repository containing Jupyter notebooks with Python implementation of each of these methods is available on [2]. The workflow is available online without any additional resource. Tables and figures are made available in the repository.

2. GEOLOGICAL CONTEXT

2.1 Unconformity-related uranium (URU) deposits

The Athabasca Basin is home to the world richest known uranium deposits. These deposits, named unconformity-related uranium (URU) deposits, are located at or in the direct vicinity (both in the sandstone and the basement) of the unconformity between an unmetamorphosed sequence of highly chemically mature, flat-lying fluvial sandstone and conglomerate forming the Athabasca Basin and its underlying Archean to PaleoProterozoic crystalline basement [3].

Uranium mineralization is structurally-controlled in relation to the presence of graphitic and/or pyritic rich conductors, and occurs mainly as semi-massive pods or veins of uraninite (pitchblende). The mineralization is spatially associated with alteration haloes composed primarily of various assemblages of clays (including illite and Mg-chlorite), tourmaline and hydrothermal hematite. Silica mobility may be intensively marked with silicified and desilicified intervals in the basement and basin lithologies.

2.2 Uranium exploration in the Athabasca Basin

Exploration for URU deposits is primarily made in two distinct steps. The first step corresponds to using airborne and ground geophysical acquisitions (EM, Resistivity) in order to delineate graphitic conductors that are thought to be spatially associated with uranium mineralization. Once favorable structures have been recognized, drilling is performed and logging occurs in the second step. As URU deposits tend to be rather limited in size, exploration is mostly targeting their more significant alteration haloes. Systematic logging of drillhole cores includes visual inspection and reporting of a variety of criteria including lithology, alteration styles in the sandstone (bleaching, diagenetic hematite, hydrothermal hematite, chlorite) and alteration and mineralization in the basement (most importantly graphite and pyrite). These logs are complemented by instrumental data, either performed during exploration campaigns (IR spectroscopy, resistivity logs, gamma ray) or after (such as routine geochemistry which is performed in the laboratory on rock samples).

3. APPLICATIONS OF MACHINE-LEARNING TO EXPLORATION DATA

Three examples of applications of ML to exploration data from the Athabasca Basin are given hereafter. Datasets related to prediction of potassic alteration and alteration facies (examples 1 and 2) were transferred by Orano Canada from its exploration projects in the eastern part of the basin. Datasets related to the prediction of lithostratigraphic units using lithogeochemistry (example 3) have been obtained from Geological Survey of Canada OpenFile 7495 [4]. Although each application of calibration or prediction problems need to be tackled specifically given the type of data under consideration, we argue that every ML workflow applied to drillhole data should at least address the following problems, which are tackled in more details in the associated repository:

— Varying resolutions between datasets. As most exploration datasets contain types of data at varying sampling intervals, one should always tackle the problem posed by varying resolutions. Some simple ways to look at

it include averaging all variables to the intervals of the lowest density variable, as well as signal-processing based solutions like wavelet tessellation [5].

— Exploratory data analysis. Understanding relationships between various variables within the dataset should always be a pre-requisite step to any ML implementation. Most popular methods include Principal Component Analysis (PCA), Cluster Analysis (CA) and Factor Analysis (FA).
— Data pre-processing. Drillhole data were often collected in remote and difficult meteorological conditions, and are thus subjected to high noise or imprecision, thus being seldom usable as such in ML routines. Good data pre-processing is thus mandatory, as most ML methods will perform better on normally distributed, smoothed datasets, which includes for geochemistry, taking into account compositional problems posed by total-sum constraints [6], handling of detection limits, or spectral pre-processing for spectra-based data [7].
— Optimization and hyperparameters tuning. Most ML methods include at least some parameters that are not learnt from data, called hyperparameters (e.g., maximum depth of trees or number of leaves in a Random Forest), and optimizing the model revolves around finding a near-optimal combination of them before evaluating the model. Various optimization approaches exist, including gridsearch based on cross-validation, randomization-based search and Bayesian optimization (the optuna library [8] is a recent implementation of this approach and was used in our workflow).
— Explainable AI. Not all ML methods are easily explainable and ML becomes valuable only as long as it is able to provide geologists with a better understanding of various processes at play in their data. It is thus important for ML models to be interpretable, i.e., being able to assess the influence of different variables in a given problem. Explainable AI is a more general problem in the AI research community and is currently an important research topic in all fields related to it [9]. When two models provide similar results, the explainable one should therefore be given preference. Decision-Tree based algorithms such as Random Forest are rather useful in this regard, as they can provide relative importance of features after prediction.

4. REAL-WORLD EXAMPLES USING EXPLORATION DATASETS FROM THE ATHABASCA BASIN

4.1 Regression of total K_2O content

4.1.1. Potassic alteration

Potassic alteration around uranium deposits occurs mainly as illite (K-rich clay), with illite haloes sometimes extending tens of meters around the orebodies, making it a first-order target for exploration. It is primarily described in logs either by visual inspection or by IR-spetroscopy. Knowledge of K_2O content of a rock can, however, be a useful feature for exploration as it gives a direct quantitative marker of the illite content of a rock (illite being the only significant K-bearing hydrothermal mineral known in the sandstone). However, geochemistry is routinely performed only on lab measurements out of exploration camps, making it unavailable during field work. On the other hand, IR spectroscopy is a fast and cheap, easily repeatable method widely available during campaigns involving very little sample preparation. Being able to infer K_2O content of a rock from IR spectroscopy could thus represent a valuable tool for geologists to guide their exploration campaigns.

4.1.2. PLS-R

Partial Least Squares – Regression (PLS-R) is a widely used method with applications ranging from medical imaging to food quality analysis and remote sensing, which is especially interesting in chemometrics [10]. It performs especially well on multicollinear data such as spectroscopy. Similarly, to Principal Component Analysis (PCA), PLS transforms a multivariate square matrix into a set of uncorrelated variables (components), with the difference that, while PCA computes components by maximizing covariance, PLS computes subsequent components by maximizing covariance with a dependent variable (the variable to predict), making it especially dedicated to prediction purposes involving high-dimensional data. Usually, optimization of PLS-R is done iteratively by running the prediction on

different successive numbers of components (1, 2… to n components) and computing Root Mean Squared Error (RMSE) and coefficient of determination R^2 on a validation dataset. For our case-study, best results were obtained using PLS-R with 22 components with RMSE = 0.02 and R^2 = 0.85.

4.2. Alteration facies prediction

4.2.1. Alteration facies in the Athabasca sandstone

Intense development of red beds at the scale of the entire Athabasca Basin is thought to have occurred shortly after deposition of the basin at 1.7Ga and is seen in most parts of the basin under the form of specular hematite along detrital quartz grains [11]. Partial to complete removal of hematite later occurred during peak diagenesis and/or hydrothermal events, yielding alternating bleached and hematized zones [12]. These red beds and their subsequent alteration are closely looked for during exploration as regional bleaching is thought to be caused by the same hydrothermal event responsible for the development of giant uranium deposits. Zones of intense bleaching and redox interfaces are thus considered good prospective targets. They are usually noted by field geologists using a visual scale ranging from none to strong. The same is done for other processes considered to be associated with uranium deposits, e.g., hydrothermal hematite, argillization, etc.

4.2.2. PLS-DA

Prior to feeding into a Machine Learning workflow, variables have been re-attributed in order to obtain more consistent classes. Two scenarios were tested, one with two classes (absence or presence of bleaching, hematite or clays), and one with three classes (trace-weak, moderate, strong). Outlier spectra were removed using the Mahalanobis distance method. As different classes had varying populations, they were randomly resampled to have the same populations, in order to correct for the sampling bias.

Partial Least Squares – Discriminant Analysis (PLS-DA) is a popular implementation of PLS-R for categorical variables. In this case, categorical variables are converted into “dummy”, binary 0 or 1 variables before running a traditional PLS. It has been fitted on the various alteration classes. For bleaching, the best results obtained gave an accuracy of 0.93 for the two-classes scenario, and 0.71 for the three-classes scenario. For diagenetic hematite, the best results gave an accuracy of 0.89 for the two-classes scenario and 0.69 for the three classes scenario.

4.3. Lithostratigraphic units

The third workflow aims at showing how to build a predictive model for lithostratigraphic units based on lithogeochemical information. OpenFile 7495 from the Geological Survey of Canada contains over 30 000 lithogeochemical analyses in Athabasca Basin [4], but only a third of them have the formation they belong to correctly labelled. As formations can have an important impact on fluid circulation and therefore on deposit formation, completing this dataset would be a valuable task and could have an importance for geological modelling in the area. The Athabasca Basin is comprised of 17 successive sub-units. Fifty samples were randomly selected from the eight main units analyzed in near-total rock analysis using ICP-MS at the Saskatchewan Research Council (Saskatoon). Best results were obtained from XGBoost classifier with an accuracy of 0.81 in attributing a sample to the right formation. The complete workflow along with implementations of various algorithms is displayed in the repository [2].

5. DISCUSSION AND CONCLUSION

Guides for implementing similar workflows are given in the associated repository along with explanations. Although not necessarily collected with applications in Machine-Learning in mind, most drillhole data used in this paper have shown to provide rather good first-order prediction results using simple algorithms. Machine-learning thus

represents a valuable opportunity to improve logging work, especially with the progressive advent of larger datasets in exploration routines (hyperspectral core loggers, core photographs analyses, XRF, etc), making an efficient use of them even more important.

ACKNOWLEDGEMENTS

Thanks are due to the French Agence Nationale de la Recherche (ANR) under grant ANR-21-CHIN-0006 (Geomin3D) and to Orano Canada for providing data from its exploration projects.

REFERENCES

[1] SCHODDE. R., "Long-term trends and outlook for uranium exploration: Are we finding enough uranium?", IAEA TECDOC SERIES, IAEA, Vienna (2018).

[2] SERDOUN. M., An Efficient Workflow for Predicting Various Logging Variables Using Machine-Learning Programs, (2023), https://github.com/MehdiSerdoun/ML_DrillingData

[3] JEFFERSON. C. W., THOMAS. D. J., GANDHI. S. S., RAMAEKERS. P., DELANEY. G., BRISBIN. D., CUTTS. C., PORTELLA. P. and OLSON. R. A., Unconformity-associated uranium deposits of the Athabasca Basin, Saskatchewan and Alberta, Bull. Geol. Surv. Canada **588** (2007) 23–67.

[4] WRIGHT. D. M., POTTER. E.G., COMEAU. J-S., Athabasca Basin Uranium Geochemistry Database, 7495, Geological Survey of Canada, Ottawa, 2014.

[5] HILL. E.J., FABRIS. A., UVAROVA, Y., TIDDY, C., Improving geological logging of drill holes using geochemical data and data analytics for mineral exploration in the Gawler Ranges, South Australia, Aust. J. Earth Sci. (2021) 1-27.

[6] AITCHISON. J., The Statistical Analysis of Compositional Data, J. R. Stat. Soc. Series. B Stat. Methodol. **44** 2, 1982, 77-139

[7] NAES. T., A User-Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, UK (2004)

[8] AKIBA. T., SANO. S., YANASE. T., OHTA. T., KOYAMA. M., "Optuna: A Next-Generation Hyperparameter Optimization Framework", Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, Anchorage (2019).

[9] VILONE. G., LONGO. L., Explainable Artificial Intelligence: A Systematic Review, (2020).

[10] MEHMOOD. T., AHMED. B., The diversity in the applications of partial least squares: an overview, J. Chemom. **30** (2015), 4-17.

[11] KOTZER. T. G., KYSER. T. K., Petrogenesis of the Proterozoic Athabasca Basin, northern Saskatchewan, Canada, and its relation to diagenesis, hydrothermal uranium mineralization and paleohydrogeology, Chem. Geol. 120 1–2 (1995) 45-89.

[12] CHU. H., CHI. G., BOSMAN. S., CARD. C., Diagenetic and geochemical studies of sandstones from drill core DV10-001 in the Athabasca basin, Canada, and implications for uranium mineralization, J. Geochem. Explor. **148** (2015) 206-230.