EXPLORE
Newsletter for the Association of Applied Geochemists
www.appliedgeochemists.org
NUMBER 193 DECEMBER 2021

Breathing new life into old assay data using machine learning methods

Tom Meuzelaar[1], Morgan Warren[1], Alice Alex[1], and Pablo Núñez Fernández[2]
[1]Life Cycle Geo, LLC., 729 Main Street, Longmont, CO, USA 80501
[2]Cobre San Rafael SL, B70491121, La Mina, S/N 15822, Touro (San Xoan), La Coruna, Spain

https://doi.org/10.70499/GCXO7847

INTRODUCTION

Significant under- or over-estimation of assay parameters can occur when incorrect laboratory assay methods are used (e.g., Abzolov 2008; Anderson 2020; Meuzelaar *et al.* 2021). The cost of re-analysis can be very high when such errors are repeated over the scale of thousands of samples. Machine learning algorithms can offer a low-cost alternative to expensive re-analysis; a small subset of samples can be re-analyzed, and an algorithm trained to: 1) recognize relationships between the corrected parameter and other assay parameters in the subset, and 2) estimate corrected values for the larger dataset.

As a proof-of-concept, machine learning algorithms were applied to 5,580 bedrock samples from the Touro exploration assay dataset to assess whether (corrected) sulfur values can be predicted from the other assay parameters in the dataset. When Atalaya Mining, Cobre San Rafael (Atalaya) acquired a majority interest in the Touro project, it inherited multiple legacy assay datasets with noticeable inconsistencies in sulfur assay data. Further investigation revealed that the data were acquired using laboratory assay methods insufficient to digest metamorphosed sulfides (predominantly pyrrhotite). Machine learning algorithms trained on a dataset with correct sulfur data were able to derive a relationship between other assay variables which enabled reproducing the sulfur concentrations with 93% accuracy. Predictive success is largely a function of: 1) the number of samples, 2) the number of assay parameters, and 3) material/deposit geochemistry.

GEOLOGICAL BACKGROUND

Proyecto Touro is a brownfield copper project located in the A Coruña province of the Galicia Autonomous Region in northwestern Spain. Copper mineralization occurs in metasediments that comprise the Órdenes Complex (Fig. 1), in the northwest portion of the Iberian Massif, an allochthonous metamorphosed unit that is part of the Variscan belt of Europe. The Órdenes Complex consists of a thick sequence of metamorphosed turbidites with interbedded volcanic lenses. These material types have undergone extensive metamorphism with sedimentary units expressing as paragneiss and volcanic units as metabasites and amphibolites. Copper mineralization occurs in the metavolcanic units as disseminated sulfides within metabasites and coarse garnet amphibolite. Sulfides are predominantly pyrrhotite and chalcopyrite, with lesser pyrite.

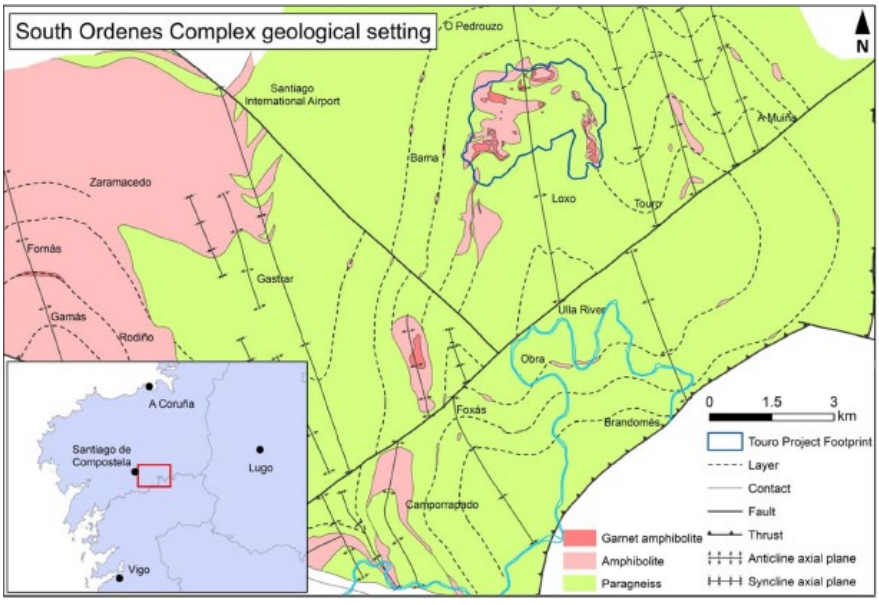

Figure. 1. Location of the Touro project area in central Spain (inset) and detailed bedrock geological map.

Touro was originally recognized as a metamorphosed Cu-Zn type volcanogenic massive sulfide (VMS) deposit (Badham and Williams 1981; Williams 1983). However, more recent studies suggest that the lithologic setting, morphology, and mineralogy more closely reflect a Besshi-type (mafic-siliciclastic) VMS deposit (Arias *et al.* 2021), equivalent to pelitic-mafic VMS deposits (Shanks *et al.* 2012). Besshi-type deposits occur in mature oceanic back-arc successions with thick marine sequences of clastic sedimentary rocks and intercalated mafic (occasionally ultramafic) volcanic rocks. The mafic component consists largely of volcanic material types with mid-ocean ridge basalt (MORB)-like affinities.

The Touro project consists of five separate mineralized zones: Arinteiro, Bama-Brandelos, Vieiro, Arca, and Monte Minas, the first three of which were mined from 1973 to 1986 (Ore Reserves Engineering 2018). Mineralization occurs

continued on page 5

Breathing new life into old assay data… *continued from page 1*

within the Arinteiro antiform, representing the final deformation episode. The antiform has NE-SW orientation with its axis plunging towards the north. The lenticular and stratiform nature of materials reflect seafloor deposition of fine-grained sediments and turbidites, intrusion of MORB basalts, and concomitant subduction/orogeny. This resulted in the current/final assemblage of mineralized amphibolites (volcanic) hosted within larger bodies of paragneiss. In Arca and Monte de Minas zones, massive sulfide-style mineralization also occurs in brecciated rocks below and near the lower amphibole contact, with breccia clasts cemented by pyrrhotite.

TOURO SULFUR DATA

Atalaya Mining acquired a majority interest in the Touro project in 2015. The acquisition included various legacy assay datasets with parameters obtained using different methods. Atalaya Mining noticed that sulfur was frequently underestimated, analyzed via multiple different lab methods, with concentrations highly inconsistent from database to database. Sulfur data are important from both an exploration and environmental perspective, as they are used in understanding ore assemblages and grade, as well as long-term waste material environmental behavior given a proportion of waste materials is likely to be acid-generating. Further investigation (Golder 2018) indicated that seven different methods for sulfur analysis had been employed over time at three different labs. Results were inconsistent between the various methods as each employed digestants of various aggressiveness, frequently resulting in partial or incomplete digestion of variably metamorphosed sulfides (Meuzelaar *et al.* 2021). Total sulfur by Leco and ICP aqua regia (Digiprep digestion) were adopted as acceptable sulfur analysis methods for the project, because results suggested near complete sulfide digestion, consistently higher sulfur assays compared to the other methods, and strong correlation between the two methods (Golder 2018).

One legacy database, with 5,880 samples and 49 assay parameters (in addition to sulfur) offered a unique opportunity to test the viability of assessing whether sulfur concentrations could be predicted from other assay parameters using machine learning algorithms. Sulfur values in this particular database were obtained by ICP-AES (four-acid) and were deemed to be of sufficient quality for this proof-of-concept evaluation as mean sulfur ICP-AES concentrations (3.4 wt. %) are 8.7% lower than the preferred Leco and Digiprep[1] concentrations (both at 3.9 wt. %). Additionally, scatterplots for these datasets indicated high coefficients of determination between both ICP-AES and Leco (0.944) and ICP-AES and Digiprep (0.955).

If successful, the methodology offers the opportunity to correct legacy datasets with misestimated parameter values without having to conduct expensive laboratory re-analysis.

METHODS

Data

A geochemical dataset from Atalaya Mining containing 5,880 samples with analysis of 49 elements was used for this evaluation. Samples for eight different bedrock lithologies were contained within the database. The sulfur values in the database ranged from below detection (<0.01) to 11.8 wt.%. A summary of the database including the lithologies, sample numbers, and median sulfur value is presented in Table 1.

Table 1. Dataset used in the study

Lithology	Code	Number of samples	Median Sulfur (wt.%)
Amphibolite	AF	1742	1.9
Garnet Amphibolite	AFG	716	3.0
Ca-poor Amphibolite	AG	1361	3.5
Breccia-Massive Sulfide	BSM	32	11.7
Biotitic Schist	DSC	297	4.3
Massive Sulfide	MS	155	5.9
Pelitic paragneiss	PG	1305	0.8
Pelitic paragneiss with sulfide	PGS	272	10.4

[1]Leco and Digiprep analyses of a subset of 97 fully digested samples were used to assess differences between ICP-AES and the total-digest analytical methods

continued on page 6

Note: *This EXPLORE article has been extracted from the original EXPLORE Newsletter. Therefore, page numbers may not be continuous and any advertisement has been masked.*

Breathing new life into old assay data… *continued from page 5*

Data Preparation

Raw geochemical data are typically not fit for advanced multivariate analysis (e.g., principal component analysis, machine learning) because of two commonly observed properties: the data are compositional in nature and may contain non-detect or censored values. Both properties of the data require mathematical transformation prior to the data being used. Compositional data present several challenges prior to statistical evaluation: 1) they are restricted to a constant sum (e.g., numeric closure), and 2) they are proportional which means when one value changes other values must also change (Grunsky and de Caritat 2020). The issue of numeric closure may be addressed with logarithmic ratios, such as the centered-log ratio (clr) (Aitchison 1982; Pawlowsky-Glahn and Egozcue 2006), which was used in this study to transform the data from concentrations to clr-transformed values. Censored data represent values below a certain analytical detection limit (DL) and are represented as being less than a value (<DL; left-censored) or greater than a value (>DL; right censored). No right-censored values were present in the dataset and left-censored values were imputed with an estimate that considers the composition of the entire sample (Sanford *et al.* 1993) using a compositional variant of the expectation-maximization (EM) algorithm (Palarea-Albaladejo and Martín-Fernández 2015). All data processing was performed in the R statistical computing environment (R Core Team 2017).

Statistical Methods

Machine learning methods can identify patterns and structures within multivariate datasets that are difficult to discern with traditional data exploration methods such as bivariate scatter plots. The objective of this study was to identify if an accurate relationship could be derived between sulfur and other elements in the database (e.g., iron, cadmium, zinc) such that sulfur concentrations could be predicted based on the composition of other elements. The relationship between sulfur and other elements is complex and highly non-linear based on the mineral stoichiometry of the lithologic units of the Touro deposit.

Multiple statistical learning methods were employed to predict the sulfur content of the Atalaya dataset. These methods included multiple linear regression, boosted decision trees (BDT; Friedman 2001), random forest (RF; Breiman 2001), and artificial neural networks (ANN; Goodfellow *et al.* 2016). Each model was built inside the Microsoft Azure Machine Learning Studio (AMLS) environment. A wide range of model complexity was chosen from simple (multiple linear regression) to advanced (ANN) to evaluate the most appropriate method. Additional complexity does not always equal additional predictive value. The relative effectiveness of any particular method is generally a function of both data density and heterogeneity.

Model accuracy was evaluated based on its ability to predict sulfur concentrations in the dataset, using both the mean squared error (MSE) and coefficient of determination (r^2). Training each model involved hyperparameter tuning of specific algorithm parameters to the dataset.

Algorithm results were interpreted by calculating the variable importance for each model by using the permutation feature-importance algorithm built in AMLS (Breiman 2001). Variable importance computes importance scores by quantifying the contribution of a specific variable (e.g., Cu, Fe, Ni etc.) on the overall model performance. In other words, variable importance computes how important each variable is in predicting sulfur.

RESULTS

Model predictive accuracy was evaluated by two metrics: mean squared error (MSE) and the coefficient of determination (r^2). Both metrics for the four models evaluated are presented in Table 2, which shows that the BDT and ANN performed significantly better than the multiple linear regression and RF models. The BDT and ANN models had very similar results, both an order of magnitude more accurate than multiple linear regression and RF. Because the BDT and ANN models performed better than multiple linear regression and RF, only those two are considered in the remaining discussion.

The predicted sulfur concentrations for the BDT and ANN are presented with the raw sulfur concentrations in Figure 2,

continued on page 8

Breathing new life into old assay data… *continued from page 6*

Table 2. Model Results

Model	MSE	r^2
Multiple Linear Regression	2.47	0.66
Random Forest	9.56	0.22
Boosted Decision Trees	0.46	0.93
Artificial Neural Network	0.43	0.93

which shows a graph of the kernel density distribution for the predictions and raw sulfur data for all 5,880 samples from the dataset. As can be seen, a good approximation for the raw sulfur distribution is generated by both models, except for the lower range of sulfur concentrations (less than 0.1 wt. %), where both models underperformed. Figure 2 illustrates that the BDT model does a slightly better job at predicting low sulfur concentrations than the ANN model.

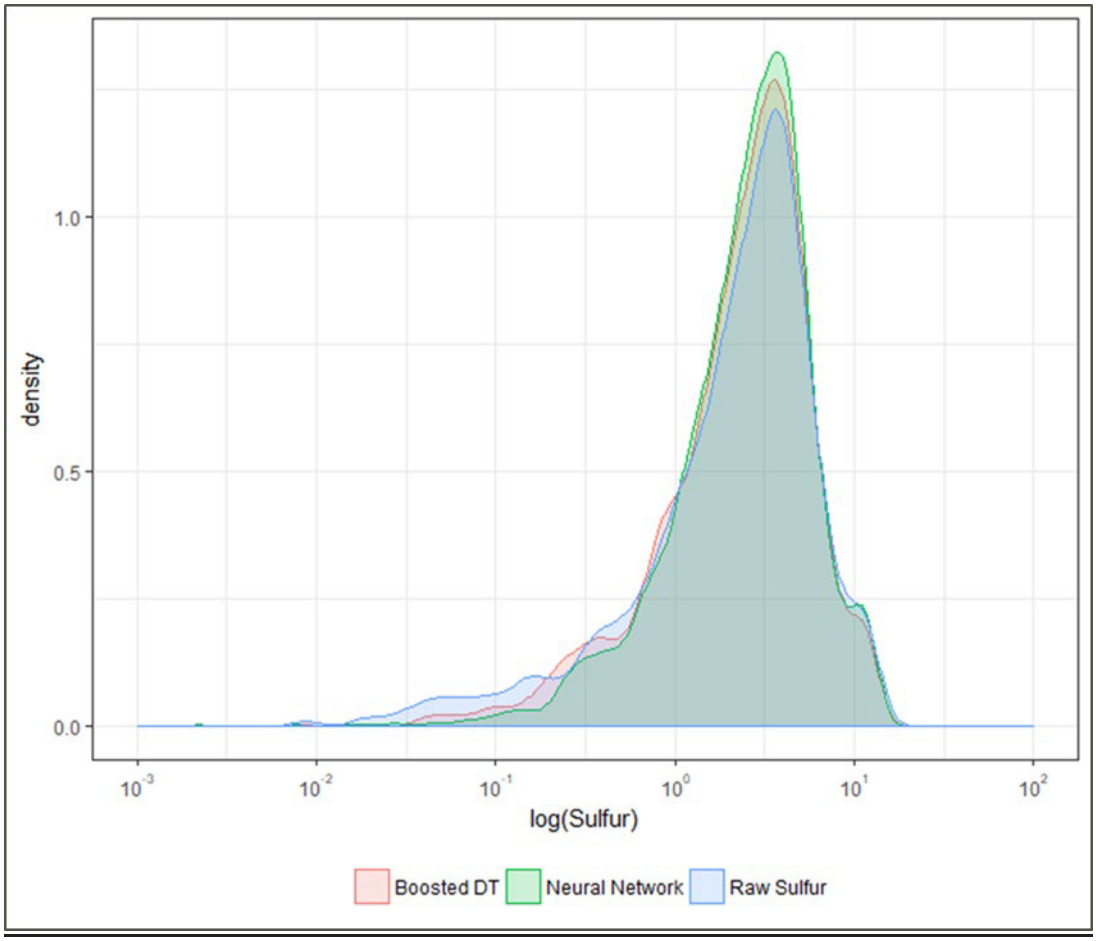

Figure 2. Distribution of predicted and raw sulfur concentrations.

The variable importance results for the top 10 ranked elements (out of 48 available) are summarized in Table 3 for the BDT and ANN models. The variables are ranked by their normalized effect on model accuracy. The top 10 elements for the BDT model are predominantly siderophile and chalcophile elements commonly associated with sulfide minerals (with Al and Na the only two exceptions). The highest ranked elements for the ANN model are more varied including elements commonly associated with sulfides (Fe and Tl), but also elements associated with carbonates (Ca, Sr, Rb, and U), and other lithophiles like Cr.

DISCUSSION

The predictive accuracy (Table 2) and variable importance (Table 3) results confirm the viability of the approach. The BDT results, in particular, are intuitive as the bulk of the elements that are important in predicting sulfur concentrations are chalcophile metals that reside in pyrite and pyrrhotite. For both methods, Fe is by far most influential which is, again,

continued on page 9

Breathing new life into old assay data… *continued from page 8*

Table 3. Variable importance rank for Boosted Decision Trees and Artificial Neural Networks

Ranking	Boosted Decision Trees		Artificial Neural Network	
	Element	Normalized Ranking	Element	Normalized Ranking
1	Fe	0.234	Fe	0.283
2	Cd	0.134	Sr	0.089
3	Zn	0.109	Th	0.069
4	Co	0.064	Ti	0.068
5	Ag	0.060	Tl	0.053
6	Cr	0.054	Cr	0.046
7	Al	0.051	Ca	0.030
8	Ni	0.034	Ba	0.030
9	Na	0.030	U	0.029
10	Se	0.025	Rb	0.026

intuitive given it is the primary component along with sulfur in primary sulfide minerals in Touro metasediments and metavolcanics. The success of the approach relies on the fact that multi-element mineral chemistry is ultimately governed by simple stoichiometric, crystallographic, and mass balance rules and is, therefore, predictable.

The fact that sulfur in this dataset is somewhat underestimated (~9% relative to Leco and Digiprep) due to incomplete digestion likely indicates that predictive accuracy would be even higher had metamorphosed pyrrhotite and pyrite been completely digested by the four-acid method. Prediction of sulfur concentrations in this dataset was intended as a proof-of-concept for the machine learning approach. Virtually all sulfur resides in two sulfide minerals of similar composition (e.g., pyrrhotite and pyrite). A simpler approach of re-analyzing sulfur for a subset of samples and using simpler regression methods based on sulfur alone was also tried and provided similar results. Hence it should be noted that a multivariate approach is not always merited when, in some cases, simpler approaches will suffice. However, in cases where parameters are distributed across multiple mineral groups (e.g., calcium in carbonates, feldspars etc.) the multivariate approach generally proves superior. The success of the approach is also dependent on sample number and geologic context. Some datasets are insufficiently small for a multivariate approach. Additionally, some datasets reflect geologic systems where parameters are distributed more randomly than others (e.g., disseminated vs. stratabound ore); such datasets will require more samples to achieve sufficiently high predictive accuracy.

Finally, multi-element geochemistry can be used to predict many other things using the machine learning approach. For example, the authors and others have used the approach successfully to predict lithology, alteration, material density, long-term environmental behavior, ore grade, metallurgical characteristics, ore vectors, and more.

ACKNOWLEDGMENTS

The authors wish to thank Atalaya Mining for providing the dataset and for adopting an innovative mindset as part of their management approach. In particular, we wish to thank Julian Sanchez, Fernando Diaz-Riopa of Atalaya, as well as Monica Barrero and Alan Noble (Atalaya contractors). We also thank Golder as much of this work was done as part of an internal innovation initiative; Mahajan Padmanathan provided significant support and guidance with IT infrastructure. Finally, we are grateful to Microsoft for providing access to the Azure cloud computing and analytics; Hubert Duan of Microsoft gave considerable and thoughtful support. Thanks to Pim van Geffen and Beth McClenaghan for their thoughtful input to this manuscript.

REFERENCES

Abzalov, M. 2008. Quality control of Assay data: A review of procedures for measuring and monitoring precision and accuracy. *Exploration and Mining Geology*, **17**, 131-144.

Aitchison, J. 1982. The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society*, Series B (Statistical Methodology), **44**, 139-177.

Anderson, S. 2020. Mount Dore Corridor: Analysis for new economy minerals. Collaborative Exploration Initiative. Chinova Resources

Arias, M., Nuñez, P., Arias, D., Gumiel, P., Castañón, C., Fuertes-Blanco, J., and Martin-Izard, A. 2021. 3D geological model of the Touro Cu deposit, a worldclass mafic-siliciclastic VMS deposit in the NW of the Iberian Peninsula. *Minerals*, **11**, 85. https://doi.org/10.3390/min11010085

continued on page 10

Breathing new life into old assay data… *continued from page 9*

Badham, J. and Williams, P. 1981. Genetic and exploration models for sulfide ores in Metaophiolites, Northwest Spain. *Economic Geology*, **76**, 2118–2127. https://doi.org/10.2113/gsecongeo.76.8.2118

Breiman, L. 2001. Random Forests. *Machine Learning*, **45**, 5-32

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **29**, 1189-1232.

Golder Associates, Inc. 2018. Sulphide and sample representativeness evaluation – Proyecto Touro, Atalaya Mining. Draft technical memorandum from T. Meuzelaar to J. Sanchez, M. Barrerero and A. Noble at Atalaya Mining, 31 p.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org

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

Meuzelaar, T., Nunez-Fernandez, P., Martin-Izard, A., Arias-Prieto, D., and Diaz-Riopa, F. 2021. The waste rock of the Touro copper deposit in Galicia, Spain: challenges for its environmental characterization. *Geochemistry: Exploration, Environment, Analysis*. https://doi.org/10.1144/geochem2020-081

Ore Reserves Engineering. 2018. Technical Report on the Mineral Resources and Reserves of the Touro Copper Project. Report prepared for Atalaya Mining Plc. 332 p.

Palarea-Albaladejo, J. and Martín-Fernández, J.A. 2015. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems, **143**, 85-96.

Pawlowsky-Glahn, V. and Egozcue, J.J. 2006. Compositional data and their analysis: an introduction, in Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (eds) Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, Special Publications, **264**, 1-10.

R Core Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Sandford, R.F., Pierson, C.T., and Crovelli, R.A. 1993. An objective replacement method for censored geochemical data. *Mathematical Geology*, **25**, 59–80.

Shanks III, W.C.P., Koski, R.A. *et al.* 2012. Volcanogenic massive sulfide occurrence model. In: Shanks III, W.C.P. and Thurston, R. (eds) *Mineral Deposit Models for Resource Assessment,* Scientific Investigations Report 2010-5070-C, U.S. Geological Survey, 1–345, https://doi.org/10.3133/sir20105070C

Williams, P.J. 1983. The genesis and metamorphism of the Arinteiro-Bama Cu deposits, Santiago de Compostela, northwestern Spain. *Economic Geology*, **78**, 1689–1700. https://doi.org/10.2113/gsecongeo.78.8.1689.