Application of Unsupervised Machine Learning Methods to Mine Water Quality Data

Tom Meuzelaar[1], Allie Wyman[1], and Shannon Zahuranec[1]

[1]*Life Cycle Geo, LLC., 729 Main Street, Longmont CO 80501, tom@lifecyclegeo.com (ORCID 0000-0003-3119-8390), allie@lifecyclegeo.com, shannon@lifecyclegeo.com*

Abstract

Mine water quality data is both ubiquitous and highly under-used. Water quality data is typically collected on a regular basis starting early in the mine project life cycle all the way into post-closure. The data is spatially extensive, multivariate and highly dimensional in nature, making it ideal for analysis by modern unsupervised methods. Application of unsupervised machine learning methods to water quality data in four different parts of the mine project life cycle (baseline and compliance, feasibility and permitting, operations and closure) at various project sites indicates that the multivariate approach succeeds where traditional methods fail and provides considerable additional insight and value to mine operators. The methods are highly effective in identifying unique water quality fingerprints, the existence of mine influence, and mixing and various reactive processes that occur in water. Example use cases are provided herein and demonstrate that the methods have been successfully used to reduce water treatment capital and operating expense, prevent and mitigate expensive environmental liability and provide significant forecasting insight.

Keywords: Mine water quality, unsupervised machine learning, life cycle water management

Introduction

Surface water, groundwater, process water and other water quality data types are collected throughout the mine project life cycle at regular, recurring intervals (e.g., MEND 2009). During early exploration stages, water quality is collected for baselining purposes. In the mine planning and stage, water quality data collected from material characterization testing is combined with the baseline data to develop models that predict water quality associated with future mining activities. Prediction outcomes support material and water management strategies, as well as permitting needs. During operations, water quality data is collected to monitor mine facility seepage, water treatment performance, and compliance at site boundaries. In closure, water quality is collected to assess closure strategy effectiveness and employ adaptive water management strategies. Water quality data is ubiquitous through the project life cycle and yet remains highly underutilized. Employment of multivariate approaches to analyzing water quality data remains the exception and not the rule.

The abundance of this data is ideally suited to using machine learning approaches to maximize its value. Water quality data is typically collected at regular intervals throughout the project life cycle. Furthermore, compliance and operational requirements dictate that water quality be collected at sufficient spatial density (MEND 2009). Finally, water quality data is multi-dimensional, commonly including measurements for 20 to 40 (or more) different parameters. The large number of parameters makes for a "wide" dataset with considerable statistical and geochemical variance, which facilitates application of innovative unsupervised machine learning methods.

In this study, unsupervised machine learning methods are applied to evaluate mine water quality during four separate stages of the mine life cycle. The names of

mine sites, operators, and other location-specific details are often not divulged as many of these analyses have yet to be reviewed by respective regulatory stakeholders. As such, the unsupervised analyses must, for now, remain in the confidential domain (except where explicitly stated).

Methods

All water quality data was analysed using Principal Component Analysis (PCA) using various Python (Van Rossum *et al.* 2009) libraries including pandas, numpy, pyrolite (Williams *et al.* 2020), sklearn, matplotlib and seaborn. Unsupervised multivariate data analysis is a typical first step in a machine learning project, primarily aimed at exploring data structure and identifying classes of related samples (often referred to as "domains"). Numerous methods exist that can be applied to water quality data (Huang *et al.* 2022) which offer various strengths and weaknesses, primarily the extent to which they preserve local and global data structures. PCA is highlighted in this study because statistical relationships between water quality parameters are easily observed on a biplot and add considerable interpretability to the results. However, it should be noted that other multidimensional methods can add considerable insight during a multivariate analysis. PACMaP (Pairwise Controlled Manifold Approximation and Projection; Wang *et al.* 2021), for instance, is recognized to be one of the most optimal methods for preserving both local and global data structure.

However, PCA is perhaps most widely used as its output can add considerable elements of explainability to the results, which is critical for machine learning work. PCA is a dimensionality reduction technique that simplifies a dataset to its irreducible, basic structure (i.e., principal components) in terms of statistical variance. Typically, the bulk of the dataset geochemical variance is accounted for by principal components 1 and 2, although for datasets with deeper statistical variance, additional components can add relevance in an analysis. The biplot is a graphic tool that displays how individual chemical elements relate to one another in principal component space. The position of chemical vectors on a biplot indicate whether elements are closely related or inversely related. Elements with longer vectors exert greater control over the overall dataset statistical variance. Individual sample factor scores can be plotted on the biplot as well which helps indicate which samples exert the greatest influence on a dataset for a particular set of parameters. For groundwater and surface water data, when sample points for a given location cluster in one area of the biplot, it indicates that water quality is not changing much over the sampling period. Locations with highly variable sample scores over time are likely undergoing some chemical changes, such as being influenced by mine impacted water, mixing with another water, or other geochemical reactions.

All water quality datasets required data cleaning and transformation prior to statistical analysis to address issues such as data censoring (detection limits), missing values, outliers and implicit numeric correlation (e.g., Aitchison, 1982).

Results

Fig. 1 provides a biplot for water quality collected for a large underground mine permitting project, where several groundwater types are expected to be encountered that might require eventual treatment during operations.

The mine will be developed in fractured bedrock (Deeper Aquifer) below a thick layer of alluvium (Shallow Aquifer). The future operator was unable to distinguish groundwater quality between the two layers using simple time series evaluation and requested a more sophisticated analysis. Fig. 1 presents results of this analysis and shows that unsupervised methods are clearly able to distinguish between deeper aquifer water, with arsenic and fluoride particularly diagnostic, and shallow waters that are more defined by their carbonate mineral composition, as indicated by the alkalinity, calcium and magnesium vectors; these trends can be directly related to logged lithology and bulk geochemical data for the boreholes. PCA results underscore the importance of dissolved trace element concentrations

in identifying each forensic signature. Furthermore, one well (Aquifer Mixing) has a very long well screen that is open to both the Shallow and Deeper aquifer – PCA clearly shows that water quality in this well exhibits both aquifer forensic signatures and makes it possible to tell during which sampling intervals the water carries one signature or the other.

Fig. 2 shows the multivariate signature of humidity cell test (HCT) data collected during the feasibility/permitting stage for the proposed Pebble mine in the USA (this mine was never permitted). This data is publicly available through a published EIS.

The PCA results displayed on the biplot are augmented by multivariate clustering (colored panels in background of Fig. 2). Multivariate analysis of HCT data has, to our knowledge, not been published and yet is highly useful. This Pebble HCT dataset was published and first presented in 2024 (Meuzelaar and Wyman, 2024) and clearly shows geochemical transitions that mine waste materials undergo in humidity cells over time in the process of becoming acid generating. The four broad clusters speak to this, with cells typically starting out at stable pH (Cluster 1), being buffered by alkalinity available in the materials. If the waste materials have negligible or low sulfide content, cell leachate quality remains in this cluster for the duration of testing. Cells with higher sulfide, however, begin to oxidize over time and enter the yellow transition zone characterized by the sulfate vector, which represents sulfide oxidation, and the magnesium, calcium and barium vectors, which represent buffering of acidity by neutralizing carbonate minerals. Once material neutralization potential is depleted, the cells start to become acid generating and enter the cluster 3 domain on the biplot – this transition is typically very fast. This cluster is defined by acidity and all

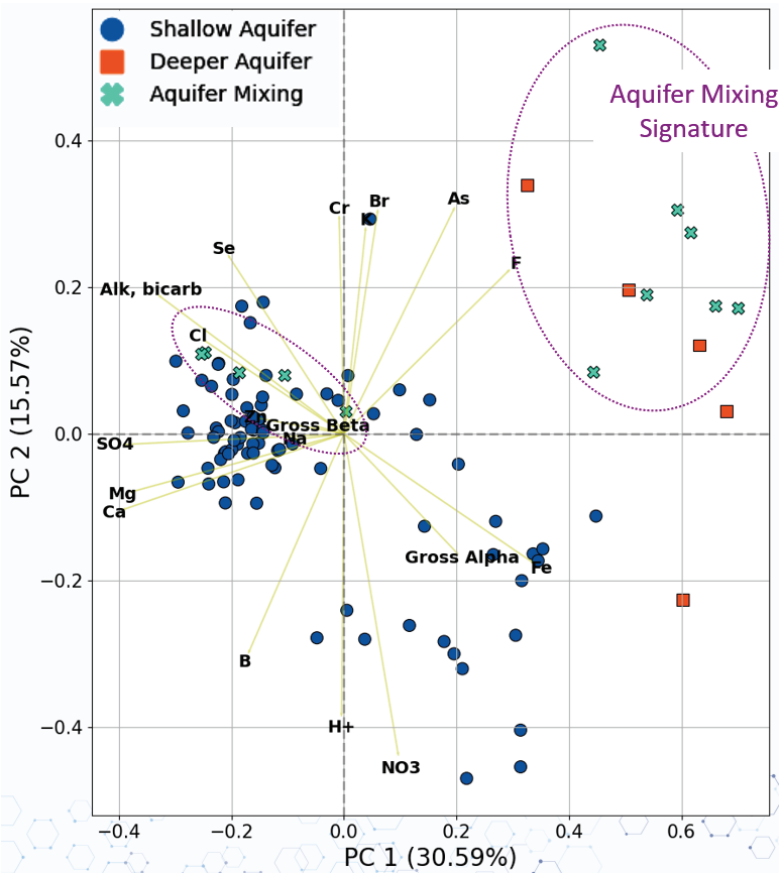

Figure 1 Water Quality Signatures of Shallow Alluvial, Deep Bedrock and a Mixed Zone.

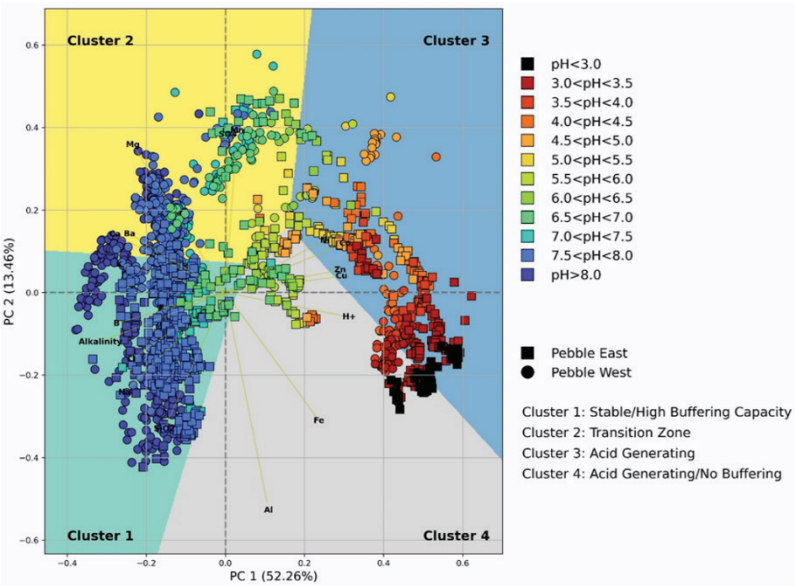

Figure 2 *Water quality signatures of the Pebble HCT dataset over time (Meuzelaar and Wyman 2024).*

the chalcophile trace metals that are typically released from sulfide minerals and remain in solution as the pH lowers. It is important to note that changes in HCT chemistry can be observed on the biplot that are often not evident based on leachate pH alone – results from unsupervised multivariate analysis can serve as an early warning indicator of ARD onset, where HCT pH often cannot. Finally, the iron and aluminum vectors (which plot towards the southeast) define cluster 4 and are consistent with a humidity cell running out of 'fuel' (sulfide content) which eventually leads to pH recovery – as pH recovers, iron precipitates out first (above pH 3–3.5), followed by aluminum (pH 4–4.5).

Fig. 3 illustrates HCT permitting data (grey symbols) collected for a confidential mine site with operational seepage data (purple symbols). The operational seepage data is overlain as the larger purple symbols in Fig. 3 and represents drainage from a lined waste rock facility consisting of materials that have long-term ARD potential. Laboratory HCT tests indicate that materials have a propensity to become acid-generating within a decade, however this has not been observed at the operational scale.

The operator is concerned that their waste rock pile will become acidic in the next few years and requested development of a seepage water quality analysis tool that will predict onset of ARD before it happens based on monthly monitoring of seepage from the waste rock pile. Early warning will give them time to respond adaptively and proactively – once ARD starts, it is extremely difficult to reverse. Because the HCTs represent accelerated weathering that generates ARD in the lab, it provides an excellent laboratory proxy for the geochemical "direction" (on the biplot) that seepage water quality is likely to move in if the waste rock pile begins to generate acid. The biplot looks somewhat similar to the previous Pebble HCT dataset, except these rocks are much more mafic in nature, so trace metals such as copper and nickel tend to predominate in leachate chemistry.

Unsupervised analysis of actual seepage data, overlain on the permitting-stage laboratory data, allows the operator to see, on a monthly basis, whether waste rock pile seepage is moving closer to acidic conditions or further away. This allows the operator to respond proactively (i.e., add lime to the pile) to prevent ARD formation which could result in costly compliance penalties.

Fig. 4 shows the final unsupervised analysis, representing four decades of groundwater and surface water quality data being collected at a mine that is going into closure.

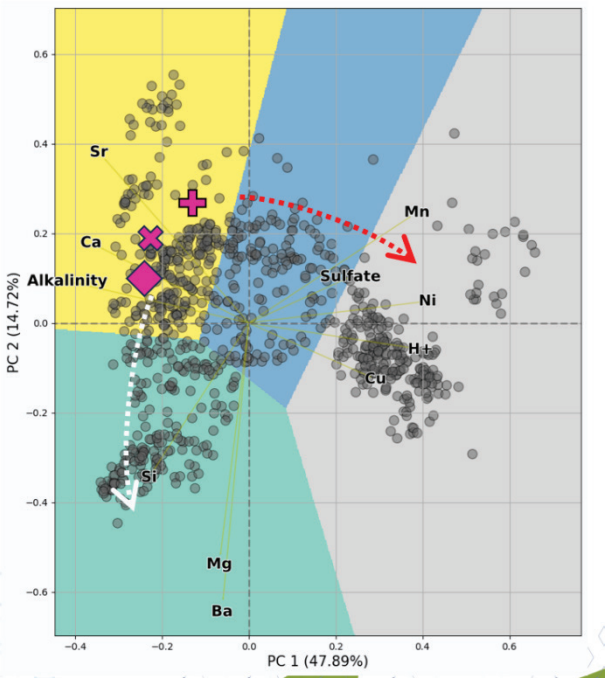

Figure 3 *Waste Pile Seepage plotted over HCT data to predict onset of ARD.*

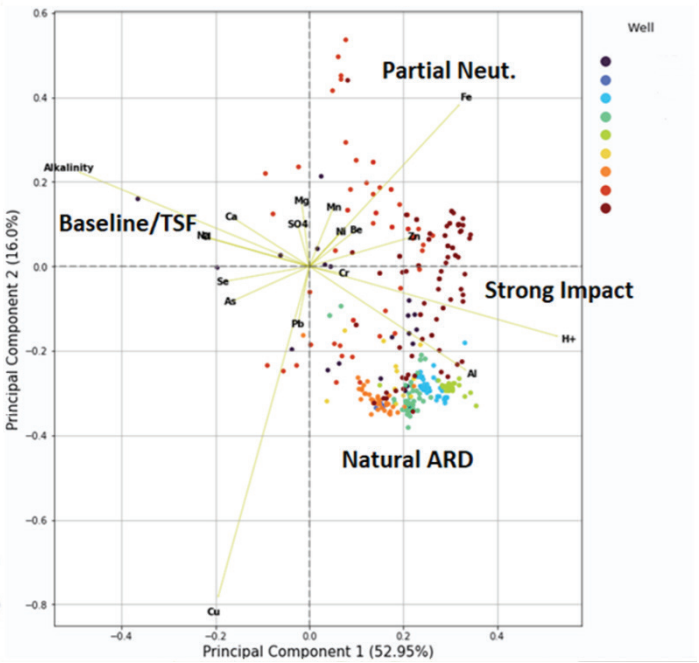

Figure 4 *PCA biplot clearly delineating natural ARD from MIW.*

The mine has known pre-mining natural ARD, as established by field observation. Large ferricrete deposits of geologic age, speak to a "fossil" acid drainage that has been partially neutralized resulting in precipitation of iron from solution, and co-precipitation of a number of metals. Once the mine went into operation, the highly reactive waste rock also began to generate acidity and MIW. Over time, the operator has been unable to separate the older, geologic waters from MIW.

Unsupervised machine learning readily differentiates natural ARD from MIW, based on the aluminum and copper signature of the former, and the fact that natural, geologic waters tended to retain highly stable water quality over the course of mining. MIW, on the other hand, displays a highly variable chemical signature over time. Pit waters (which plot towards the right and northeast corners of the biplot) especially, are highly variable over the sampling period. This is consistent gradual formation of ARD and declining pH over time, as indicated by the proton and base metal vectors on the biplot. Unsupervised analysis is also able to differentiate MIW of different types, as tailings seepage and background groundwater are defined more by alkaline, acid-buffering (calcium) and cation exchange (sodium) processes. Further unsupervised analysis of just the tailings and baseline groundwater dataset (not shown here) yields further success in delineating tailings MIW from natural baseline groundwater; the latter often has high sulfate and chlorite due to strong evaporative ambient conditions and is again not easy to discriminate from tailings MIW using traditional, lower-dimensional analysis methods.

Separation of natural ARD from pit-generated MIW indicates that all MIW is generated within the pit hydraulic capture zone. Additional analysis of area water quality signatures further indicate that waste rock dumps are not generating seepage or impacts. The fact that all impacts are hydrologically contained changed the project closure strategy significantly as the initially perceived need for future active water treatment is now likely eliminated. This is a very significant savings for the client, that was made possible by using more sophisticated data analysis methods.

Conclusions

Application of unsupervised methods to various water quality data types collected in four different parts of the mine project life cycle indicates that the multivariate approach is highly effective in identifying different water quality domains, breakthrough of MIW, mixing effects and various reactive processes that occur in water.

The four use cases have application and implications as follows:

- **Baseline and compliance:** machine learning methods are applied to separate pre-mining water quality in different hydrostratigraphic units and to detect vertical transmissivity between units, providing the future operator considerable flexibility and power in future water treatment planning, and also providing significant insight to the site hydrogeologic conceptual model.
- **Feasibility and permitting:** unsupervised analysis of a large laboratory HCT dataset provides considerable insight to the nature and timing of acid conditions; this information is highly valuable in making HCT termination decisions as lag time to acidity can be challenging to predict using standard methods (depletion calcs, monitoring changes in leachate pH).
- **Operations:** comparison of waste rock seepage leachate to HCT data generated during permitting allows for regular monitoring of seepage quality and provides early warning to potential acidic conditions, allowing the operator to react and respond proactively.
- **Closure planning:** multivariate analysis of decades of groundwater and surface water quality resulted in successful delineation of natural ARD signatures from various types of MIW. The ability to identify these water types influences the long-term closure strategy and is likely to result in considerable water treatment cost savings.

Acknowledgements

The authors wish to thank its confidential clients for allowing use of their datasets for this type of complex analysis and for being forward thinking and open to these more sophisticated approaches. We also thank our LCG colleagues Alice Alex, Morgan Warren and Sam Wright for assisting with various aspects of data analysis and review.

References

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

Huang, H., Yingfan, W., Rudin, C., Browne, E. 2022. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. Communications Biology, v. 5, no 719

MEND. 2009. Prediction Manual for Drainage Chemistry from Sulphidic Geologic Materials. MEND Report 1.20.1. December 2009.

Van Rossum, G., Drake, F. L. 2009. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace. https://docs.python.org/3/reference/.

Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y. 2021. Understanding how Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualation

Williams *et al*. 2020. pyrolite: Python for geochemistry. Journal of Open Source Software, 5(50), 2314. https://doi.org/10.21105/joss.02314

Wyman-Feravich, A., Meuzelaar, T. 2024. The Spiral toward Acidity: A Machine Learning Approach to Analyzing Humidity Cell Test Results, International Conference for Acid Rock Drainage (ICARD 2024), Halifax Canada.