Goldschmidt 2025 Abstract
https://doi.org/10.7185/gold2025.26867

Advancing Geochemical Understanding through Machine Learning Classification of Large Spatio-Temporal Soil and Groundwater Data

MEHDI GHARASOO[1], SILVIO JANETZ[1], MARCUS FAHLE[1], TAMARA VAN STADEN[2], PHILIPPE VAN CAPPELLEN[2] AND CHRISTOPH NEUKUM[1]

[1]Federal Institute for Geosciences and Natural Resources (BGR)
[2]University of Waterloo

Advances in field measurement techniques have enabled rapid data acquisition while simultaneously facilitating the collection of long-term, spatially extensive datasets, resulting in increasingly large and complex data inventories. These datasets often challenge traditional statistical methods in deriving meaningful insights. However, with the growth of machine learning and deep learning techniques, the potential for analyzing such data has improved significantly. Yet, the variety of the techniques, each tailored to specific needs, presents a challenge in selecting the most appropriate method and preprocessing the raw data accordingly. This process often confounds researchers at the outset.

In this study, we present two case studies involving distinct spatio-temporal datasets. The first dataset consists of phosphorus concentrations in soils across all counties in southern Ontario over a 30-year period. This dataset is continuous and comprehensive, containing about two hundred thousand data points, focusing only on a single compound. The second dataset comprises groundwater quality analyses (e.g., iron, sulfate, nitrogen compounds, etc.) collected at random points and intervals over several decades in a mining-impacted region of Eastern Germany. Unlike the first dataset, this data is sporadic and scattered, with missing values due to specific, targeted campaigns focused on particular measurements at certain times and locations.

We investigate the application of classification techniques, such as k-means and hierarchical clustering, for interpreting these datasets and propose methods to simplify the data, making it suitable for modeling with geochemical tools such as PHREEQC. The results of this approach contribute to a deeper understanding of anthropogenic drivers of geochemical changes across various geological structures and scales, thereby enhancing our conceptual understanding of groundwater-soil dynamics.