Data Before Algorithms: An Open-Source Approach to Legacy Geochemical Method Categorization

Sam Scher[1], Tom Carmichael[2]
[1]LKI Consulting Inc, Washington, United States, [2]Datarock, Adelaide, Australia

Legacy geochemical datasets frequently contain ambiguous analytical methods, undocumented modifications, and inconsistent reporting standards. These issues significantly impede data integration, interpretation, and, critically, the confidence placed in geochemical analyses by non-geochemist stakeholders, such as data scientists and machine learning specialists. The growing trend of multidisciplinary use of geochemical data has increased the risk of inappropriate inclusion or exclusion of these datasets due to misinterpretation or perceived complexity.

Currently, addressing analytical ambiguity predominantly relies on manual approaches, including subjective leveling and detailed univariate analyses, methods that are labor-intensive, inefficient, and potentially error-prone. To enhance objectivity and streamline this process, we developed a generalized, reproducible workflow implemented through open-source R code designed to automatically categorize instrumental geochemical data based solely on statistical distributions and patterns of missingness inherent to the datasets.

This workflow starts with exploratory data analysis to systematically detect distinct data subsets by evaluating precision, accuracy thresholds, and joint probability density functions. By identifying characteristic data distributions and missing data signatures, our methodology minimizes reliance on subjective judgment. However, acknowledging the inherent complexity of geochemical data, a final manual step of leveraging probability plots and spatial distributions ensures the robustness of method categorization.

Our approach addresses a critical technological gap by ensuring that interpretations derived from legacy datasets reflect accurate analytical contexts. By providing clarity to analytical origins, we mitigate common pitfalls associated with multidisciplinary usage of geochemical data, ultimately preventing flawed interpretations and enhancing the reliability of data-driven mineral exploration decisions. Thus, this method places appropriate emphasis on robust data characterization before applying advanced analytical technologies, avoiding scenarios where the technological cart precedes the geochemical horse.