Insight into Geological Processes in Deep Time: Exploratory Multivariate Analysis of Geochemical Datasets

Morgan Williams[1,2] and Jens Klump[1]

[1] CSIRO Mineral Resources, [2] Australian National University

MINERAL RESOURCES
www.csiro.au

Rocks and minerals can retain geochemical records of their origin and history, which can be utilised to provide otherwise inaccessible constraints on the evolution of our planet. We are working with existing databases and developing an analysis framework to address data complexities and provide robust constraints for a series of key problems in geochemistry. Here we highlight the principal data-related challenges, and enumerate the major components of the analysis workflow.

From Data, Geology

Much has happened since our planet was a primitive ball of molten rock, including the origin of plate tectonics, the modern atmosphere and life. This extended geological history has been encoded into chemical signatures of rocks and minerals, which may then used to (partially) reconstruct the past.

Inverting geochemistry to infer the geological past is commonly an underdetermined problem (especially prior to the advent of modern geochemical analysis instrumentation), and is hindered by complex geological histories.

Modern analytical methods have higher throughput and greater sensitivity and precision. As a result, established publicly-accessible geochemical databases are growing steadily[1]. However, the potential value of aggregating the increasing volume of high-quality data has not yet been fully realised.

Confined Spaces and Missing Values

Geochemical data is compositional in nature (i.e. sums to 100%), and statistical analysis requires appropriate log-transformations[2] (Fig. 1). These transformations are sensitive to null- and below-detection values, and in practice some form of parametric imputation is required in order to perform data analysis[3,4].

Global geochemical databases include analyses of varying quality and provenance conducted over several decades. Issues regarding compositional data are compounded by high-dimensionality and low overall data density.

Aggregation of multiple incomplete sample records increases their overall value and minimises null-values, but due to compositional relationships between parameters, aggregation becomes a complex standardisation task.

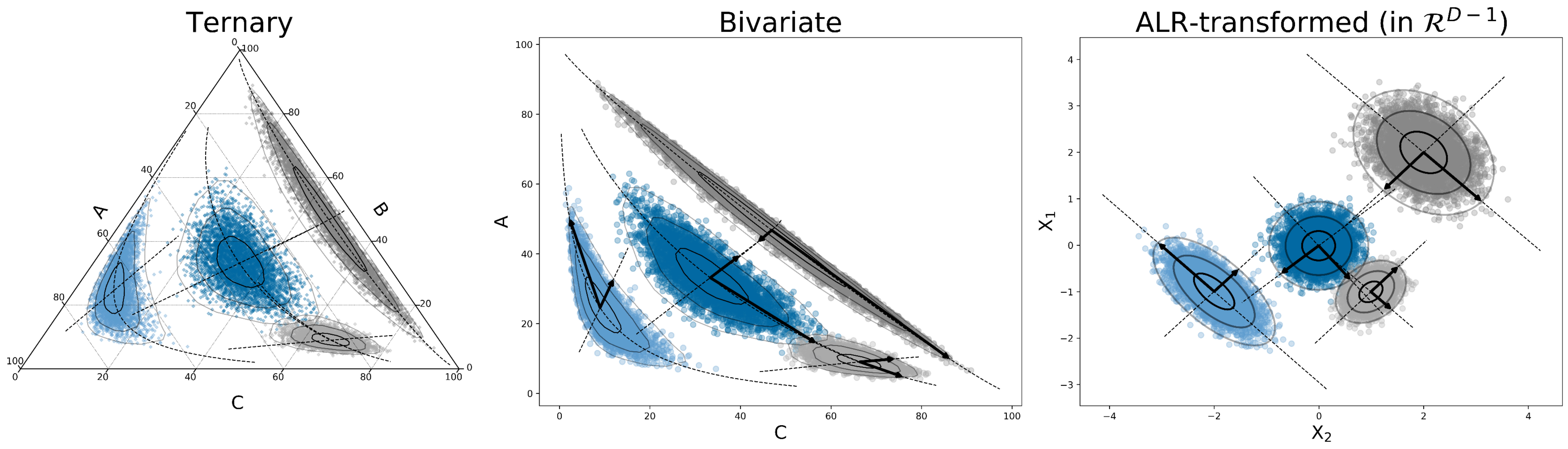


Figure 1: Synthetic data illustrating the log-normal distributions of compositional random variables, and their equivalents as commonly visualised by geoscience professionals (ternary, bivariate plots), and after log-transformation (right).

Data Analysis Workflow

Once a number of data issues have been addressed, appropriate transformations allow statistical quantification of compositions and their relationships. Here we attempt to include as much information for interpretation and inference, and incorporate additional data-derived features based on domain knowledge (e.g. geochemical proxies, rock and mineral classification schemes) and specific reference frames (e.g. reference compositions, global ranges).

Collate

- Data aggregation
- ‘Fill the gaps’: Complement databases with additional data
- Schemas, ontologies

Clean

- Formatting
- Data validation
- Recalculation and ‘unification’

Filter

- On metadata
- On data quality
- Alteration proxies

Transform

- Data reduction
- Add proxy features
- Log-transforms
- Imputation
- Dimensional reduction (e.g. PCA)

Vizualize

- Ternary, bivariate, log-transformed
- Biplots

Model

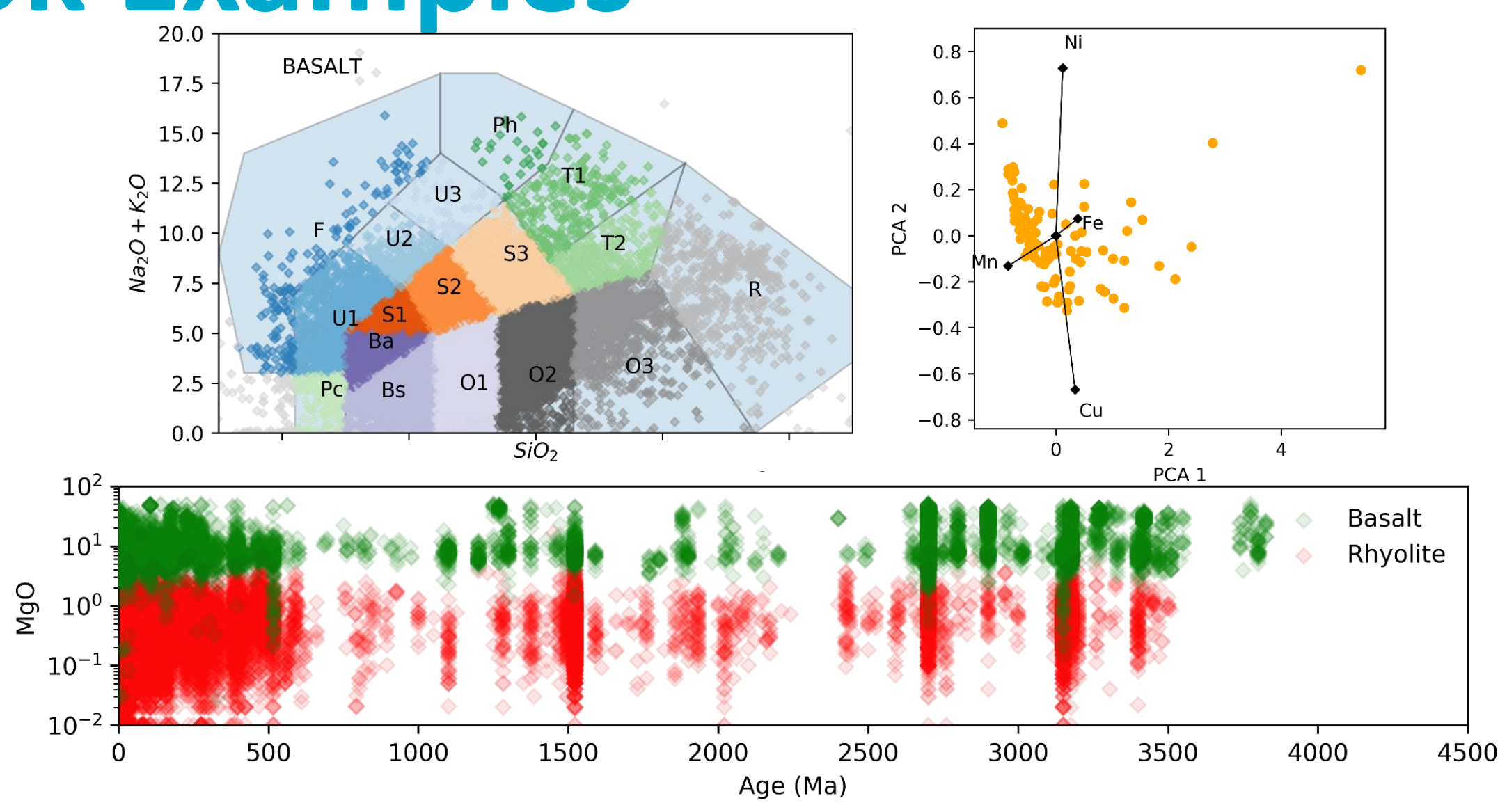
- Physical/chemical process models
- Mixture models
- Clustering, networks, semantic relationships

Test

- Hypothesis testing
- Assess sampling and geological bias
- Sensitivity testing (incl. to imputation)

Jupyter Notebook Examples

github.com/morganjwilliams/exploratory-geochemistry



Prospects and Continuing Work

- Geochemical data is ‘not so small, but complex’.
- Multivariate analysis will aid detection of detect subtle contrasts.
- Improving integration of domain-specific knowledge, including semantics.
- Future focus on process modelling and use of probabilistic reasoning.

FOR FURTHER INFORMATION

Morgan Williams
e morgan.williams@csiro.au
w https://www.csiro.au/en/Research/MRF
github.com/morganjwilliams/exploratory-geochemistry

REFERENCES

[1] Lehnert, K. et al. (2000). A global geochemical database structure for rocks. Geochemistry, Geophysics, Geosystems 1. [2] Aitchison, J. (1982). The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society. Series B 44, 139–177. [3] Palarea-Albaladejo, J. and Martín-Fernández, J.A. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computers & Geosciences 34, 902–917. [4] Martín-Fernández, J.A. and Thió-Henestrosa, S. (2006). Rounded zeros: some practical aspects for compositional data. Geological Society, London, Special Publications 264, 191–201.

Where relevant, bibliographic information is included in code documentation.

DATA SOURCES

The data used in this project is predominantly sourced from the EarthChem database, which is part of the Interdisciplinary Earth Data Alliance (IEDA) Data Facility and hosted at the Lamont-Doherty Earth Observatory (Columbia University).

ACKNOWLEDGEMENTS

This work builds upon a large body of work by geologists and statisticians, and John Aitchison in particular. The components of this workflow are being developed in Python and use a variety of open-source packages and tools. Online examples are presented using Jupyter notebooks hosted free on Binder and built using the Dash framework by Plotly.