Tephra compositional data: are we doing it right?

Simon Larsson[1] & Matthew Bolton[2]

[1]Department of Physical Geography, Stockholm University, Stockholm, Sweden (simon.larsson@natgeo.su.se)
[2]Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Canada (bolton1@ualberta.ca)

1. Background

Tephra (volcanic ashes) used for geochronological applications are identified based on geographic and stratigraphic context, glass shard morphology, and, perhaps most important for differentiation of samples, geochemical composition. The latter is most commonly analysed by electron probe microanalyser (EPMA) and presented as weight percentages of the nine or ten most abundant elements, often normalised to a 100% total for ease of comparison. Simple exploration of such data with reference to previous findings usually allows confident identifications, but new findings continuously add to the complexity with an increased likelihood of multiple candidates for identification.

2. Rationale

As datasets become increasingly complex, statistical analyses are needed more frequently. Tephrochronologists often use principal component and discriminant function analyses, but rarely consider that compositional data suffer from the constant-sum constraint and require log-ratio transformations for mathematically sound analyses. There is yet to be a consensus on a tephra compositional data curation procedure that includes log-ratio transformations. To address this issue, we are exploring and comparing various approaches to compositional data treatment to determine whether a formal recommendation for such a procedure is relevant for the tephra community.

3. Method

EPMA data is collected from published literature
to assemble datasets of different sizes and complexities

↓

Datasets are normalised to a 100% sum total

↓

SMOTE is used to equalise class imbalances

↓

Datasets are treated to cope with zero values:

- Imputation (artificial replacement of zeroes)
- Column removal (exclusion of elements with zeros)
- Row removal (exclusion of samples with zeroes)

↓

Variations of log-ratio transformation are performed

↓

Principal Component Analysis is performed
on both normalised data and log-ratios

↓

Classifiers are trained on resulting datasets:

- LDA (Linear Discriminant Analysis)
- NB (Naïve Bayes, with kernel density generative model)

↓

Performance is measured by repeated 10-fold cross-validation

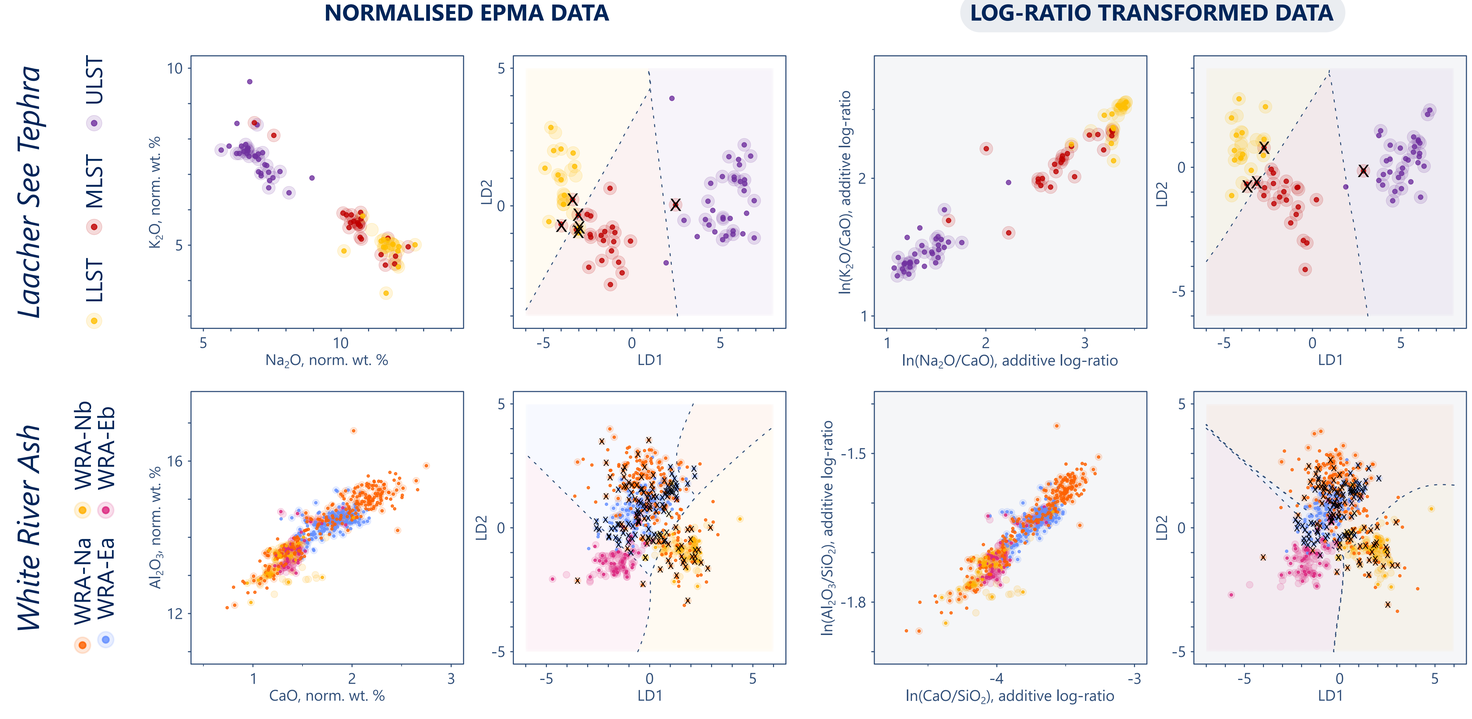
↓

Variable importance is evaluated using ROC curve metrics
(receiver operating characteristic, for binary or multi-class classification)

↓

Results are considered to determine if there is an optimal data treatment process that yields the most reliable separation between related tephras

4. Preliminary results



Tephra classification accuracy may be slightly improved with log-ratio transforms (but not always)!

References for data displayed in this poster

Harms & Schmincke (2000) Volatile composition of the phonolitic Laacher See magma (12,900 yr BP): implications for syn-eruptive degassing of S, F, Cl and H_2O. *Contributions to Mineralogy and Petrology*, 138, 84-98.
Preece et al. (1999) Tephrochronology of late Cenozoic loess at Fairbanks, central Alaska. *GSA Bulletin*, 111(1), 71-90.
Preece et al. (2014) Chemical complexity and source of the White River Ash, Alaska and Yukon. *Geosphere*, 10(5), 1020-1042.
Tomlinson et al. (2020) Chemical zoning and open system processes in the Laacher See magmatic system. *Contributions to Mineralogy and Petrology*, 175, art.no. 19.

Financing of the presenting author's participation at EGU23 gratefully received from the Swedish Research Council and Albert & Maria Bergström's stipend fund.