Proposed Short Course Title: Geological Applications of Compositional Data Analysis: A Practical Introduction

Mark Engle[1] and Madalyn Blondes[2]
[1]Eastern Energy Resources Science Center, U.S. Geological Survey, El Paso, TX
[2]Eastern Energy Resources Science Center, U.S. Geological Survey, Reston, VA

Many types of data used in the earth sciences are compositional, meaning they are composed of relative variables or parts. Common examples include soil texture or grain size distribution, elemental or oxide concentrations, relative abundance of minerals or coal macerals, etc. It has been shown that, for mathematical reasons, analysis of compositional data using classical tools, such as correlation or even simple plotting, can provide results that are misleading or incorrect. Methods based on log-ratio transformations, so-called Compositional Data Analysis (CoDA), have been developed in an attempt to prevent some of these known problems in the interpretation and analysis of compositional data. In this short course, we will introduce attendees to the basic concepts of CoDA including the 3 basic log-ratio transformations, run through real-world geological examples using the CoDaPack software package. The latter portion of the course will cover some more advanced topics and current areas of research and provide citations for future information.

Proposed outline/schedule (1-full day):

Morning (9AM-12PM)

1) Introduction to compositional data
    a. Defining compositional data
        i. Class discussion: what are and are not compositional data?
    b. Examples of known problems with compositional data
        i. Spurious correlations and regressions
        ii. Subcompositional Incoherence (interpretations change based on which subset of elements are used)
2) Basic concepts of compositional data analysis (CoDA)
    a. Transformations from constrained Aitchison Space (The Simplex) to unconstrained Euclidean Space
    b. Brief history of log-ratio development
        i. Which log-ratio transformations address which problems of closed data?
3) Univariate and bivariate analysis
    a. Describe approaches
    b. Class Exercise & Discussion: Real data example using CoDaPack

Lunch Break (12PM-1PM)

Afternoon (1PM-4PM)

4) Multivariate exploratory data analysis
    a. Introduction to the centered log-ratio (clr)

    b. Interpretation of a clr biplot
    c. Class exercise and discussion: example using CoDaPack
5) Creation and uses of the isometric log-ratio (ilr)
    a. Class exercise: Creation of singular binary partition using food example
    b. Examples using stoichiometric principles and from literature
6) Advanced topics
    a. Missing and censored data
    b. Modeling variables
    c. Incorporation of non-compositional variables (e.g., depth and temperature) into multivariate CoDa

About the instructors: Dr. Blondes and Dr. Engle have authored and co-authored several peer-reviewed journal articles and a book chapter incorporating CoDA into geochemical interpretation of various earth systems. They are active in the CoDA community, both making presentations at CoDA workshops in Europe. In addition, Dr. Engle has taught the basics of CoDA to students as part of his co-taught geochemical data analysis course at the University of Texas at El Paso.

Relevant publications by the instructors:

Shelton, J. L., Engle, M. A., Buccianti, A., & Blondes, M. S. (2018). The isometric log-ratio (ilr)-ion plot: A proposed alternative to the Piper diagram. *Journal of Geochemical Exploration*, *190*, 130–141.

Blondes, M. S., Engle, M. A., & Geboy, N. J. (2016). A practical guide to the use of major elements, trace elements, and isotopes in compositional data analysis: Applications for deep formation brine geochemistry. In J. A. Martín-Fernández & S. Thió-Henestrosa (Eds.), **Compositional Data Analysis** (pp. 13–29).

Engle, M. A., & Rowan, E. L. (2014). Geochemical evolution of produced waters from hydraulic fracturing of the Marcellus Shale, northern Appalachian Basin: A multivariate compositional data analysis approach. *International Journal of Coal Geology*, *126*, 45–56.

Engle, M. A., & Blondes, M. S. (2014). Linking compositional data analysis with thermodynamic geochemical modeling: Oilfield brines from the Permian Basin, USA. *Journal of Geochemical Exploration*, *141*, 61–70.

Engle, M. A., & Rowan, E. L. (2013). Interpretation of Na-Cl-Br systematics in sedimentary basin brines: Comparison of concentration, element ratio, and isometric log-ratio approaches. *Mathematical Geosciences*, *45*, 87–101.