Estimates in compositional data analysis: is it possible to use original units?

J.A. Martín-Fernández[1]

[1]University of Girona, Girona, Spain; *josepantoni.martin@udg.edu*

Abstract

Compositional Data (CoDa) have been defined historically as random vectors with strictly positive components whose sum is constant (e.g., 100, one, a million). The term CoDa has been expanded to include other units such as mol per liter, microgram per m^3, euros, minutes, pieces, animals or tones. That is, the term CoDa covers all vectors representing parts of a whole which only carry relative information, being an equivalence class in their mathematical representation (Barceló-Vidal and Martín-Fernández 2016). Nowadays working in orthonormal log-ratio coordinates, one can apply any multivariate technique to estimate values for CoDa. However, difficulties appear when the composition does not have a constant sum and the analyst wants the estimates expressed in their original units. This is the case when working with compositions measured, for example, in euros, in mol per liter or in milligram per m^3, and in the most common case when the analyst deals with non-closed compositions expressed as vectors of proportions. This scenario is shared in a number of scientific areas such as geochemistry, economy, medicine or ecology. Because for recovering the original units of the estimates one must add an extra variable (column) to the original dataset, in this contribution we explore two sound approaches:

1. To add to all the compositions a residual with respect to an arbitrary closure constant. The orthonormal basis is accordingly extended by one to obtain the log-ratio coordinates for the statistical analysis.
2. To perform the statistical analysis working in a T-space. That is, consider the orthonormal coordinate representation of the vector formed by the log-score of the geometrical mean of the parts and the log-ratio coordinates of the composition.

The invariance of both approaches under a change of orthonormal basis is explored. In addition, using simple data sets, the properties of both approaches are compared. Importantly, for the first approach, the concern is whether the original units of the estimates depend on the arbitrary constant and, if so, what the behavior is when the constant approaches infinity.

Key words: Concentrations, Logratio, Percentages, Proportions.

References

Barceló-Vidal, C. and Martín-Fernández, J.A. (2016). The Mathematics of Compositional Analysis. newblock *Austrian Journal of Statistics 45*(4), pp. 57–71.