

Compositional data and the geometry of its sample space

Egozcue, J. J.¹, Pawlowsky-Glahn, V.²

SUMMARY

Compositions are arrays with positive components and we are interested in the information contained in the ratios between the components. Frequently, a composition can be viewed as describing the parts of some whole and are represented as proportions, percentages, concentrations, shares, etc. Standard statistical methods fail at analysing raw compositional data as they are affected by spurious correlation. The problems with standard statistical methods point out that they are originated in the inadequacy of the geometry of the underlying sample space. Nowadays, compositional data analysis relies in that the simplex, as sample space, is endowed with the Aitchison geometry, which is a particular Euclidean geometry: the operations are perturbation and powering, which play the role of addition and scalar multiplication; the inner product is based on log-ratios of the components. The main characteristic of this geometry is that all functionals relevant to the analysis are scale-invariant, i.e. are 0-degree homogeneous functions. As in any Euclidean space, the elements can be represented by Cartesian coordinates known as isometric log-ratio coordinates. The consequence is that all statistical methods designed for multivariate real variables can be safely applied on samples in isometric coordinates.

Keywords: Compositional data

AMS Classification: -

¹Department of Ingeniería Civil y Ambiental
Universidad Politécnica de Cataluña
email: juan.jose.egozcue@upc.edu

²University of Girona
email: vera.pawlowsky@udg.edu