Paper Number: 5557

From constant sum constraint to association of compositional variables

Egozcue, J.J.[1]
[1]Universitat Politècnica de Catalunya, Campus Nord, Edif. B2, E-08034 Barcelona, Spain,
juan.jose.egozcue@upc.edu

Compositional data were characterised in the sixties as data which components add to a fixed constant, like 1 (proportions), 100 (percentages), or a million (ppm) [1]. For this reason they were called closed data. This characterisation induced the false idea that, suppressing a component or using units like mg per litre, micrograms per m3 or molar concentrations, problems with spurious correlation disappear and that the data were no longer compositional. The log-ratio approach (Aitchison) put forward the fact that the information provided by compositional data is coded in the ratios between the parts or components independently of the constant sum constraint. Since 2000, compositional data are defined as equivalence classes of vectors with proportional components [2,3] which are represented in the simplex, endowed with a particular geometry called Aitchison geometry [4]. It is characterised by the use of coordinates (log-contrasts) in order to represent compositions. All these ideas are the core of what we call the sample space approach to compositional data analysis.

Along this process of profiling concepts and structure of compositional data, applied scientists received the message that statistical correlation is something that cannot be used with compositional data thus making the use of statistics cumbersome. However, the sample space approach provides new ways of studying association between parts of a composition. Proportionality of parts across a compositional sample gives an alternative to correlation [5]. This association is generalizable to groups of parts within a composition.

A very simple example of salinization of ground water chemistry allows discussing the previous concepts: the data do not add to a constant, a subcomposition is analysed, spurious correlation is dramatically present, coordinates do not depend on the used units, but association of Cl and Na is still detectable using the compositional analysis standard tools [3].

References:

[1] Chayes, F. (1960), J. Geophys. Res., 65(12), 4185–4193
[2] Martín-Fernández J A, C Barceló-Vidal and V Pawlowsky-Glahn (2003), Math. Geol., 35(3), 253-278
[3] Pawlowsky-Glahn V, J J Egozcue and R Tolosana-Delgado (2015), *Modelling and Analysis of Compositional Data*. John Wiley & Sons, 272pp
[4] Pawlowsky-Glahn, V and J J Egozcue (2001), SERRA, 15(5), 384-398
[5] Lovell D, V Pawlowsky-Glahn, J J Egozcue, S Marguerat and J Bähler (2015), PLoS Comput. Biol., 11(3)