Compositional data: simple questions, difficult answers; or implications of the sample space choice

J. J. Egozcue[1] **, and V. Pawlowsky-Glahn**[2]
[1]Universidad Politècnica de Catalunya, Barcelona, Spain; *juan.jose.egozcue@upc.edu*
[2] Universitat de Girona, Spain

Abstract

Early definitions of compositional data were based on the constant sum of the components. In the eighties, John Aitchison complemented this definition with some properties and principles. However, a formal definition of compositional data and their different typologies is still pendent. Frequently, although not free of controversial opinions, compositional data are identified with those data that are analysed with the log-ratio approach. This implies that the attention is directed to the ratios between components. However, there are cases in which the parts do not have the adequate scale, e.g. the scale can be closer to the absolute scale more than to the ratio scale required for a log-ratio approach. Here two main points are addressed: (I) the scale of the data is frequently hidden by the constant sum constraint or other characteristics of the data, and (II) when data are claimed to be parts of a whole, there is no indication about whether these parts are overlapping or not.

Some simple questions may illustrate the lack of precision of the present definitions. Let us formulate one of them.
A Professor examined her students two times and their assessments consist of a score per exam. The scores are numbers from 0 to 10, both included. In order to study the relation between the two scores the Pearson correlation coefficient is computed. Is this correlation coefficient an appropriate measure of dependence? Is it spurious? Should a 4 score be considered as the pair $(4, 6)$? Is this a Likert scale? Are these data compositional? If yes, can they be treated using log-ratios, or may be should they be identified with a clr-transformed composition?

The first step of any statistical analysis should be to decide which is an appropriate sample space for the available data and which is its mathematical structure (operations, scales, distances, projections). This structure should allow answering the stated questions. However, a given data set may be viewed within different sample spaces with different structures,and each structure has different implications. The adequacy of the choice is referred to answers obtained for the stated questions. The definition of any type of compositional data requires a detailed description of the sample space and its structure.

Key words: scale, spurious correlation, overlapping categories, compositional equivalence, Aitchison geometry