Towards a pragmatic approach to compositional data analysis

Michael Greenacre
Universitat Pompeu Fabra, Barcelona, Spain; *michael.greenacre@upf.edu*

Abstract

In this talk I will return to first principles of compositional data analysis, as originated by John Aitchison (1986), who used ratios of parts as the fundamental starting point for description and modelling. I show that a compositional data set can be effectively replaced by a set of ratios, one less than the number of parts, and that these ratios describe an acyclic connected graph of all the parts. Contrary to recent literature, I show that a particular set of ratios based on the additive log-ratio transformation, proposed initially by Aitchison, can be an excellent substitute for the original data set, as shown in an archaeological data set as well as in three other examples. Whereas it has been recommended to avoid the additive log-ratio transformation because it "deforms" the log-ratio distances, these examples show that this transformation loses very little of the distance geometry and is adequate for all practical purposes. The advantage of additive log-ratios is their simplicity of interpretation, whereas other log-ratio transformations that have been proposed, such as centred and isometric log-ratios, have nice theoretical properties but no intuitive univariate interpretation for a practitioner.

I propose further that, using any pairs of components (or "parts"), a smaller set of ratios can be determined, either by automatic selection, expert selection or a combination of both, which explains as much of the total log-ratio variance as required for all practical purposes. These component ratios can then be validly summarized and analyzed by conventional univariate methods, as well as by multivariate methods, where the ratios are preferably log-transformed.

In summary, this work points towards a simpler and more pragmatic approach to compositional data analysis, leading to log-ratios that are easily interpretable, comparable across studies, and measurably close to the more complex methods proposed up to now. This approach is similar in spirit to that proposed by Greenacre (2011), who defined a measure of subcompositional incoherence, and then investigated methods that were departing slightly from perfect coherence that performed in practice just as well as the strictly coherent ones.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman & Hall, London. Reprinted in 2003 with additional material by Blackburn Press.

Greenacre, M. (2011). Measuring subcompositional incoherence. *Mathematical Geosciences* **43**, 681–693.