THE RESEARCH GROUP ON STATISTICAL ANALYSIS OF COMPOSITIONAL DATA

Carles Barceló
Departament d'Informàtica i Matemàtica Aplicada
Universitat de Girona

Compositional data analysis deals with positive, relative data, which express proportions of some whole. Typical examples are data recorded as percentages, ppm, ppb, g/kg, or in relative units, such as data in atomic or molecular proportions which are obtained as the ratio of weight percent to atomic or molecular weight. Compositional data are common in all fields of applied science, from natural to social science.

Historically, only the first type of data has been considered to be compositional, and the fact that percentages are naturally subject to a unit-sum constraint has been considered the special and intrinsic feature of this type of data. In applications, this unit-sum constraint has been widely ignored or wished away, and inappropriate 'standard' statistical methods, devised for and successfully applied to unconstrained data, have been used with disastrous consequences. The solution was given in the early 1980's by John Aitchison, who introduced a new methodology based on the additive logratio transformation from the d-dimensional simplex sample space to d-dimensional real space. This approach was extended to spatially dependent compositional data (Pawlowsky-Glahn, 1984, 1989; Pawlowsky-Glahn and Burger, 1992; Barceló-Vidal, and Pawlowsky-Glahn, 1995), to finite mixtures of compositions (Barceló-Vidal and Pawlowsky-Glahn, 1995; Barceló-Vidal, Pawlowsky-Glahn and Grunsky, 1995), and to the detection of outliers in compositional data (Barceló-Vidal, Pawlowsky-Glahn and Grunsky, 1996). Presently, a group of researchers from the Universitat de Girona, the Universitat Politècnica de Catalunya in Barcelona, and the Universitat de Barcelona –together with John Aitchison, professor emeritus of the University of Glasgow, and other researchers from the Free University of Berlin (Germany), the Universities of Jena (Germany), Firenze (Italy), and Kansas (USA), as well as the Canadian Geological Survey– are working on further developments of this methodology, emphasizing the geometric aspects of compositional data analysis and the applications to specific problems in geology and archaeometrics.

Geometric approach to statistical analysis on the simplex: According to Aitchison, perturbation and power transformation induce a real vector space structure in the simplex. Furthermore, an inner product, with associated norm and distance can be defined. The resulting Euclidean space structure of the simplex creates a framework within which all statistical methods devised for data in real space can be integrated. In particular, previous concepts specific to compositional data appear as the natural counterpart to equivalent concepts in real space, like measures of location or measures of variability (Pawlowsky-Glahn and Egozcue, 2001*a*, 2001*b*).

The natural sample space of compositional data: The recent developments of the vector space structure of the simplex have led to a more extensive study of the mathematical rationale underlying compositional data. This study has led to the definition of the natural sample space of compositional data as a quotient space, a datum thus being a class of equivalence, and reducing the simplex to one possible representative of this quotient space. This approach implies that equivalent representations are also compositional in nature, showing that the constant-sum constraint is not an inherent property to compositional data, while the relative nature certainly is. (Barceló-Vidal *et al.*, 2001).

Cluster analysis on the simplex: The compositional distance introduced by Aitchison and a new measure of difference for compositional data based on measures of divergence have been used to develop an appropriate clustering strategy for compositional data (Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn, 1997, 1998*a*, 1998*b*, 1999). Within this framework, a new multiplicative zero replacement algorithm has been proposed (Martín-Fernández Barceló-Vidal and Pawlowsky-Glahn, 2000).

Applications: Compositional data analysis is present in different scientific fields, for example in geology, archaeology, sociology and econometrics. The new methodology has been applied to solve specific problems of these fields giving rise to new concepts –coherent with the nature of the sample space– such as differential perturbation processes (Aitchison and Thomas, 1998; Buccianti, Pawlowsky-Glahn, Barceló-Vidal and Jarauta-Bragulat ,1999), compositional centering (Martín-Fernández, Bren, Barceló-Vidal and Pawlowsky-Glahn, 1999; Eynatten, Pawlowsky-Glahn and Egozcue, 2002), and others (Pawlowsky-Glahn and Barceló-Vidal, 1999; Martín-Fernández, Olea-Meneses and Pawlowsky-Glahn, 2001; Pawlowsky-Glahn and Buccianti, 2001). Sometimes the compositional methodology has been misunderstood by the specialists of these fields and therefore supplementary efforts have been devoted to convince the scientific community of the coherence and advantages of the approach (Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 1999; Aitchison, Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 2000, Aitchison, Barceló-Vidal and Pawlowsky-Glahn, 2001).

New probability distributions on the simplex: The multivariate skewnormal distribution introduced by Azzalini and Dalla Valle (1996), and Azzalini and Capitanio (1999) allowed to define a new parametric class of distributions on the simplex, the additive logistic skewnormal class of distributions (Mateu-Figueras, Barceló-Vidal and Pawlowsky-Glahn, 1998). This class includes the additive logistic normal distribution defined originally by Aitchison and expands the possibilities to capture the patterns of variability observed in practice on compositional data sets. The geometric approach to statistical analysis of compositional data mentioned before offers a new framework within which a systematic approach to distributions on the simplex can be undertaken.

References

1. Aitchison, J., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2001). «Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio analysis». *Archaeometry*, (in press).
2. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2000). «Logratio analysis and compositional distance». *Mathematical Geology*, 32, 271-275.
3. Aitchison, J. and Thomas, C. W. (1998). «Differential perturbation processes: a tool for the study of compositional processes». In A. Buccianti, G. Nardi and R. Potenza (eds.), *Proceedings of IAMG'98–The Fourth Annual Conference of the International Association for Mathematical Geology*. De Frede Editore, Napoli (I), 499-504.
4. Azzalini, A. and Dalla Valle, A. (1996). «The multivariate skew-normal distribution». *Biometrika*, 83, 715-726.
5. Azzalini, A. and Capitanio, A. (1999). «Statistical applications of the multivariate skew-normal distribution». *Journal of Royal Statistical Society, series B*, 61, 579-602.
6. Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1995). «Finite mixtures of compositional data». *Science de la Terre*, 32, 29-48.
7. Barceló-Vidal, C., Pawlowsky-Glahn, V. and Grunsky, E. (1995). «Classification problems of samples of finite mixtures of compositions». *Mathematical Geology*, 27, 129-148.
8. Barceló-Vidal, C., Pawlowsky-Glahn, V. and Grunsky, E. (1996). «Some aspects of transformations of compositional data and the identification of outliers». *Mathematical Geology*, 28, 501-518.
9. Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (1999). «Comment on "Singularity and Nonnormality in the Classification of Compositional Data" by Bohling, G. C. *et al.*», *Mathematical Geology*, 31, 581-586.
10. Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2001). «Mathematical foundations for compositional data analysis». *Proceedings of IAMG'01 –The Annual Conference of the International Association for Mathematical Geology*. CD-Rom. Cancún (México), 20 pp.
11. Buccianti, A., Pawlowsky-Glahn, V., Barceló-Vidal, C. and Jarauta-Bragulat, E. (1999). «Visualization and modeling of natural trends in ternary diagrams: a geochemical case study». In S. J. Lippard, A. Næss and E. Sinding-Larsen R. (eds.), *Proceedings of IAMG'99–The Fifth Annual Conference of the International Association for Mathematical Geology*. Tapir, Trondheim (N), 139-144.
12. Eynatten, H. V., Pawlowsky-Glahn, V. and Egozcue, J. J. (2002). «Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams». *Mathematical Geology*, (accepted for publication).

13. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1997). «Different classifications of the Darss Sill data set based on mixture models for compositional data». In V. Pawlowsky-Glahn (ed.), *Proceedings of IAMG'97–The Third Annual Conference of the International Association for Mathematical Geology*. International Center for Numerical Methods in Engineering (CIMNE), Barcelona, 151-158.
14. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998*a*). «Measures of difference for compositional data and hierarchical clustering methods». In A. Buccianti, G. Nardi and R. Potenza (eds.), *Proceedings of IAMG'98–The Fourth Annual Conference of the International Association for Mathematical Geology*. De Frede Editore, Napoli, 526-531.
15. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998*b*). «A critical approach to non-parametric classification of compositional data». In A. Rizzi, M. Vichi and H. H. Bock (eds.), *Advances in Data Science and Classification. IFCS-98–Proceedings of the 6th Conference of the International Federation of Classification Societies*. Springer-Verlag, 49-56.
16. Martín-Fernández, J. A., Bren, M., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1999). «A measure of difference for compositional data based on measures of divergence». In S. J. Lippard, A. Næss and E. Sinding-Larsen R. (eds.), *Proceedings of IAMG'99–The Fifth Annual Conference of the International Association for Mathematical Geology*. Tapir, Trondheim, 211-215.
17. Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2000). «Zero Replacement in Compositional Data Sets». In: H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen and M. Shader, *Proceedings of IFCS'2000, The 7th Conference of the International Federation of Classification Societies*, Namur, 155-160.
18. Martín-Fernández, J. A., Olea-Meneses, R. and Pawlowsky-Glahn, V. (2001). «Criteria to compare estimation methods of regionalized compositions». *Mathematical Geology*, (in press).
19. Mateu-Figueras, G., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998). «Modeling compositional data with multivariate skew-normal distributions». In A. Buccianti, G. Nardi and R. Potenza (eds.), *Proceedings of IAMG'98–The Fourth Annual Conference of the International Association for Mathematical Geology*. De Frede Editore, Napoli, 532-537.
20. Pawlowsky-Glahn, V. (1984). «On spurious spatial covariance between variables of constant sum». *Science de la Terre*, 21, 107-113.
21. Pawlowsky-Glahn, V. (1989). «Cokriging of regionalized compositions». *Mathematical Geology*, 21, 513-521.
22. Pawlowsky-Glahn, V. and Barceló-Vidal, C. (1999). «Confidence regions in ternary diagrams». *Terra Nostra*, 99, 37-47.
23. Pawlowsky-Glahn, V. and Buccianti, A. (2001). «Visualization and modeling of subpopulations of compositional data: statistical methods illustrated by means of

geochemical data from fumarolic fluids». *International Journal of Earth Sciences*, (in press).

24. Pawlowsky-Glahn, V. and Buger, H. (1992). «Spatial structure analysis of regionalized compositions». *Mathematical Geology*, 24, 675-691.
25. Pawlowsky-Glahn, V. and Egozcue, J. J. (2001*a*). «About BLU estimators and compositional data». *Mathematical Geology*, (in press).
26. Pawlowsky-Glahn, V. and Egozcue, J. J. (2001*b*). «Geometric approach to statistical analysis on the simplex». *Stochastic Environmental Research and Risk Assessment*, 15, 384-398.
27. Pawlowsky, V., Olea, R. A. and Davis, J. C. (1995). «Estimation of regionalized compositions: a comparison of three methods». *Mathematical Geology*, 27, 105-148.