**Vera Pawlowsky-Glahn[1] and Juan José Egozcue[2]**

[1]*Dept. of Computer Science and Applied Mathematics;*
*University of Girona; Girona, SPAIN; vera.pawlowsky@udg.edu;*
[2]*Dept. of Applied Mathematics; Technical University of Catalonia;*
*Barcelona, SPAIN; juan.jose.egozcue@udg.edu*

# THE CLOSURE PROBLEM: ONE HUNDRED YEARS OF DEBATE

## Abstract

*Back in 1897, Karl Pearson published a paper, which title began with the words "On a form of spurious correlation : : :". He was the first to point out the dangers which may befall the analyst when using conventional statistical methods with compositional data, known at that time as closed data. Many have been the scientists that have tried to understand, explain, and solve the problem since than, specially among geologists. But it was not until the 1980's that a solution was proposed, which has developed in a completely new methodology. This approach was the log-ratio approach, put forward by John Aitchison. In this contribution, we summarise the problems and the state-of-the-art in the new developments.*

## 1 Introduction

Historically, compositional data were defined as data where the elements are non-negative and sum to unity. Nowadays, compositional data are defined as data which only carry relative information. This latter definition is much broader, as it includes data which do not sum to a constant, like molar or molal compositions. Compositional data often arise from non-negative data (such as counts, area, volume, weights, expenditures) and are represented as data scaled by the total of the components, i.e. data subject to a constant sum constraint. Representing compositional data with a constant sum or as data which do not sum to a constant, is a simple change which does not alter the character of the data and which, in particular, does not modify the relative character of the information carried by them. The natural sample space for compositional data with $D$ components is the regular $D$-part simplex, $S^D$, and nowadays efforts concentrate in studying its geometry and its consequences in applications.

As stated by Aitchison (2003) and Aitchison and Egozcue (2005), the statistical analysis of compositional data has gone through roughly four phases. The first three were mainly concentrated in the XXth century, while the fourth can be considered to be part of the present state-of-the-art. They are summarized below.

## 2 Compositional data analysis in the XX$^{th}$ century

The first phase lasted approximately from 1897 to 1960. It was initiated by a paper on spurious correlations by Karl Pearson (1897). He was the first to point out the danger of interpretation of standard (Pearson) correlation of ratios of data. Standard multivariate statistical analysis, as developed at the beginning of the XX$^{th}$ century, is an appropriate form of analysis for the investigation of problems with real sample spaces. There, the usual geometric concepts, calculus and, in general, mathematical operations taught at school, can be applied without problems. But a compositional vector, subject to a constant-sum constraint, is different from an unconstrained vector. Still, scientists and statisticians insisted over several decades in studying all the pitfalls of standard multivariate analysis, in particular correlation analysis, when applied to compositional vectors. The assertion, made by John Bacon-Shone (2011), that

> *the key question is whether standard multivariate analysis, which assumes that the sample space is RD, is appropriate for data from this restricted sample space (the simplex) and if not, what is the appropriate analysis*

states the problem clearly. This question can naturally be extended to any multivariate data which have a constraint sample space and, to our understanding, the answer in all these cases is to apply the *principle of working on coordinates* (Mateu-Figueras et al., 2011).

Pearson (1897) identified the problem of "*spurious correlation"* between ratios of variables, showing that if $X$, $Y$ and $Z$ are uncorrelated, then $X=Z$ and $Y=Z$ will not be uncorrelated. Chayes (1960) linked Pearson's work and compositional data. He showed that the unit sum constraint induces negative correlations between components of the composition. However, he was unable to suggest a way to remove the effect of the constraint (Bacon-Shone, 2011). He initiated the second phase, the specific condemnation of using standard multivariate techniques with compositional data. In the second phase, which ranged roughly from 1960 to 1980, the geologist Felix Chayes was the primary critic of the application of standard multivariate analysis to compositional data. His main criticism was in the interpretation of product-moment correlation between components of a geochemical composition. He introduced the concept of negative bias, or closure problem, as the constant sum constraint forces negative correlations. Correlations of compositional data are not free to range between -1 and +1.
(Sarmanov and Vistelius, 1959) supplemented the Chayes criticism in geological applications and (Mosimann, 1962) drew the attention of biologists to it. However, distortion of standard multivariate techniques, due in particular to spurious correlation, when applied to compositional data was the main goal of study and no alternative and appropriate methodology was found.

The third phase started with the realization by Aitchison in the 1980s that compositions provide information about relative, not absolute, values of components. He stated that, to acknowledge the fact that information is relative, any reasonable statement about a composition has to be in terms of ratios of components (Aitchison, 1981a,b, 1982, 1983, 1984b,a). The fact that log-ratios are easier to be handled mathematically than ratios, and that a logratio transformation provides a one-to-one mapping onto a real space, led him to take logs of the ratios. The logratio transformation approach was born. These transformations allowed the use of standard (unconstrained) multivariate statistics applied to transformed data. Inferences could be translated back into the simplex, leading to compositional statements.

The key techniques in the third phase have been very popular and successful over more than a century; starting with the introduction of the logarithmic transformation for positive data by Galton (1879) and McAlister (1879), through variance-stabilizing transformations, the Box-Cox transformation (Box and Cox, 1964) and implied transformations in generalized linear modelling. The logratio transformation principle is based on the fact that information carried by compositions is relative and not the absolute one, and that there is a one-to-one correspondence between compositional vectors and associated logratio vectors. Any statement about compositions can be reformulated in terms of log-ratios, and vice versa. The transformation removes the problem of a constrained sample space, the unit simplex, and projects the data into an unconstrained space, multivariate real space. Original transformations were principally the additive logratio transformation (Aitchison, 1986, p. 113) and the centered logratio transformation

(Aitchison, 1986, p. 79). The logratio transformation methodology seemed to be accepted by the statistical community (see for example the discussion of Aitchison (1982)). However, the actual impact in applied sciences was, and still is, limited.

## 3 Compositional data analysis at the beginning of the XXIst century

The fourth phase arises from the realization that the internal simplicial operation of perturbation, the external operation of powering, and the simplicial metric introduced by John Aitchison in the 80's, define an Euclidean or finite dimensional Hilbert space (Billheimer et al., 1997, 2001; Pawlowsky-Glahn and Egozcue, 2001). Many compositional problems can be investigated with this specific algebraic-geometric structure, which has led to the stay-in-the-simplex approach (Mateu-Figueras, 2003; Pawlowsky-Glahn, 2003). This staying-in-the-simplex point of view proposes to represent compositions by their coordinates (Mateu-Figueras et al., 2011), as they live in an Euclidean space, and to interpret them and their relationships from their representation in the simplex. Accordingly, the sample space of random compositions is identified to be the simplex with a simplicial metric and measure, different from the usual Euclidean metric and Lebesgue measure in real space.

Two main principles of compositional data analysis are scale invariance and subcompositional coherence. Scale invariance obeys the intuitive idea that a composition provides information only about relative values not about absolute values. Therefore, ratios of components are the relevant entities to study. This concept is equivalent to the statement that all meaningful functions of a composition should be expressed in terms of ratios (Aitchison, 1997, 2002).

Subcompositional coherence demands that two scientists, one using full compositions and the other using subcompositions, should make the same inference about relations within the common parts. Ratios within a subcomposition are equal to the corresponding ratios within the full composition. Subcompositions of compositions are the analog of marginals or sub-vectors in unconstrained analysis (Aitchison, 1986, p. 33).

These principles, formulated by J. Aitchison in the eighties of the past century, find its formal counterpart in the simplex geometry, called Aitchison geometry. The main ideas can be summarized in: (a) proportional real vectors, with positive components, are equivalent; equivalence classes are compositions which can be represented by the constant sum vector (Aitchison, 1992; Barceló-Vidal et al., 2001; Egozcue et al., 2011). (b) Subcompositions can be expressed as orthogonal projections in the context of Aitchison geometry of the simplex (Egozcue and Pawlowsky-Glahn, 2005b). Point (a) is a formalization of the scale-invariance principle, while point (b) guarantees the requirements of subcompositional coherence.

The simplex of $D$ parts, $S^D$, includes all positive real vectors adding up to a given constant. Absolute values of components in a composition are meaningless unless they are compared by ratios with other components. We use the notation $S^D$, where the superscript is the number of parts of the composition. However, this superscript has been also used to indicate the dimension of the space, being $D$-1.

Basic operations in the simplex are closure, perturbation and powering.

Closure is a normalisation to a given constant $\kappa$ and consists of selection of a representative of the equivalent vector of positive components. This constant is usually unity, percentage, ppm, or ppb. It does not affect the ratios between components, and $\kappa$ is therefore unimportant.

Perturbation (Aitchison, 1986, p. 42) is computed multiplying compositions component by component and, afterwards, normalizing to the closure constant.

Perturbation has a neutral element, which is a composition with equal components. After closure these components are $\kappa /D$ in a $D$-part simplex. Any composition perturbed by this neutral element remains unaltered. The inverse operation of perturbation, is merely dividing components of a composition by the corresponding components of the other composition; closure reduces the result to an element of the simplex.

Perturbation in the simplex is analogous to translation in real space; it is a way to record change. Perturbation plays an important role also in describing imprecision, in the definition and computation of residual compositions in regression, and in other fitting techniques. From the mathematical point of view, perturbation is an Abelian group operation in the simplex.

There is a second operation in the simplex, powering. It is the analog of scalar multiplication in real space and consists of raising each component to the constant and then applying closure to the result. The operations *perturbation* and *powering* define a $D$-1 dimensional vector or linear space structure on $S^D$ (Pawlowsky-Glahn and Egozcue, 2001).

The structure can be extended to produce a metric vector space by the introduction of the simplicial metric or distance defined in Aitchison (1983) The distance is permutation and perturbation invariant, and the effect of powering is analogous to the effect of scalar multiplication in real spaces. It has also subcompositional dominance (Aitchison, 1992). This constitutes $S^D$ as a metric space. To measure angles between vectors in a metric space, an inner product is needed. Such an inner product, consistent with this metric, has been defined (Billheimer et al., 1997, 2001; Pawlowsky-Glahn and Egozcue, 2001, 2002; Egozcue et al., 2003). Together with the associated norm, the Euclidean structure of the simplex is obtained. We refer to this as the finite dimensional Hilbert space structure of the simplex, in order to distinguish it from the ordinary Euclidean structure of real spaces, and to the corresponding geometry in the simplex as *Aitchison geometry*, to distinguish it from the ordinary Euclidean geometry of real spaces. But they have completely analogous properties.

As for any vector space, generating vectors, bases, linear dependence, orthonormal bases, and subspaces play a fundamental role. For instance, alr coordinates are not orthonormal, and many problems associated with this transformation are due to not taking this fact into account; the clr transformation corresponds to a generating system, which explains why the covariance matrix of clr coordinates is singular.

Orthonormal bases are important because they provide a straightforward way of computing the coefficients or coordinates of a composition. The coefficients of a $D$-part composition $\mathbf{x}$ relative to an orthonormal basis, can be computed as the inner product of $\mathbf{x}$ with the elements of the orthonormal basis. They are called coordinates with respect to that basis. Coordinates are log-ratios. They are called isometric log-ratios (ilr) since they preserve the simplicial metric in $S^D$ (Egozcue et al., 2003). The transformation that assigns the coordinates to the composition **x** allows the computation of distances, norms, and inner products, as ordinary Euclidean ones when using the coordinate vectors. Within the ilr framework we can get different transformations corresponding to different orthonormal bases. A very intuitive way is based on a sequential binary partition (SBP) (Egozcue and Pawlowsky-Glahn, 2005b, 2006b). The approach leads to balances and to a graphical representation, called balance-dendrogram (Egozcue and Pawlowsky-Glahn, 2005a, 2006a; Pawlowsky-Glahn and Egozcue, 2011), which is very helpful for interpretation. Coordinates are by definition orthogonal logcontrasts (Aitchison, 1986, p.85), involving ratios of compositional components in a more complicated way than simple log-ratios and so may pose more difficult problems in interpretation. Selection of adequate orthonormal bases plays a central role for data analysis.

## 4 Conclusions

We think that the interesting future of compositional data analysis is twofold. On the one hand, it will lie in statisticians searching for real applied problems. Applications reveal that there is still a long way to depurate statistical methods applied to compositional data and the interpretation of results. For instance, the development of robust statistical methods (Filzmoser and Hron, 2011) or the treatment of zeroes (Martín-Fernández et al., 2011). On the other hand, the recent extension of the compositional approach to infinite-dimensional spaces (Egozcue et al., 2006; van den Boogaart et al., 2010) has opened up a whole field of theoretical problems that can be tackled with this approach. We share the idea of Tchebycheff, expressed in his Theory of Maps, *Real progress is made when theory and the needs of application go hand in hand*. The state-of-the-art in this field of research can be found in a book (Pawlowsky-Glahn and Buccianti, 2011) to honour John Aitchison at his 85th birthday.

## References

Aitchison, J. (1981a). Distributions on the simplex for the analysis of neutrality. In *Statistical Distributions in Scientific Work-Models, Structures,and Characterizations*, pp. 147-156.

Aitchison, J. (1981b). A new approach to null correlations of proportions. *Mathematical Geology 13* (2), 175-189.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44* (2), 139-177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika 70* (1), 57-65.

Aitchison, J. (1984a). Reducing the dimensionality of compositional data sets. *Mathematical Geology 16* (6), 617-636.

Aitchison, J. (1984b). The statistical analysis of geochemical compositions. *Mathematical Geology 16* (6), 531-564.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London (UK): Chapman and Hall Ltd. (Reprinted 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology 24* (4), 365-379.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97 - The III Annual Conference of the International Association for Mathematical Geology*, Volume I, II and addendum, Barcelona (E), pp. 3-35. International Center for Numerical Methods in Engineering (CIMNE).

Aitchison, J. (2002). Simplicial inference. In M. A. Viana and D. S. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics (American Mathematical Society)*, pp. 1-22. University of Notre Dame, Notre Dame, Indiana: American Mathematical Society, Providence, R.I. Aitchison, J. (2003). Compositional data analysis: where are we and where should we be heading? See Thió-Henestrosa and Martín-Fernández (2003). CD-ROM.

Aitchison, J. and J. J. Egozcue (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology 37* (7), 829-850.

Bacon-Shone, J. (2011). Compositional data analysis: an historical review. See Pawlowsky-Glahn and Buccianti (2011), pp. 3-11. 378 p.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 - The VII Annual Conference of the International Association for Mathematical Geology*, Cancun (Mex), pp. 20 p. Kansas Geological Survey.

Billheimer, D., P. Guttorp, and W. Fagan (1997). Statistical analysis and interpretation of discrete compositional data. Technical report, NRCSE technical report 11, University of Washington, Seattle (USA), 48 p.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96* (456), 1205-1214.

Box, G. E. P. and D. R. Cox (1964). The analysis of transformations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 26*(2), 211-252.

Chayes, F. (1960). On correlation between variables of constant sum. *Journal of Geophysical Research 65* (12), 4185-4193.

Egozcue, J. J., C. Barceló-Vidal, J. A. Martín-Fernández, E. Jarauta-Bragulat, J.-L. Díaz-Barrero, and G. Mateu-Figueras (2011). Elements of simplicial linear algebra and geometry. See Pawlowsky-Glahn and Buccianti (2011), pp. 141-157. 378 p.

Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawlowsky-Glahn (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica 22* (1).

Egozcue, J. J. and V. Pawlowsky-Glahn (2005a). Coda-dendrogram: a new exploratory tool. In G. Mateu-Figueras and C. Barceló-Vidal (Eds.), *Proceedings of CoDaWork'05, The 2nd Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona, ISBN 84-8458-222-1, http://ima.udg.es/Activitats/CoDaWork05/.

Egozcue, J. J. and V. Pawlowsky-Glahn (2005b). Groups of parts and their balances in compositional data analysis. *Mathematical Geology 37* (7), 795-828.

Egozcue, J. J. and V. Pawlowsky-Glahn (2006a). Exploring compositional data with the coda-dendrogram. In E. Pirard, A. Dassargues, and H. B. Havenith (Eds.), *Proceedings of IAMG'06 - The XI Annual Conference of the International Association for Mathematical Geology*, Liµege (B). University of Liège, Belgium, CD-ROM.

Egozcue, J. J. and V. Pawlowsky-Glahn (2006b). Simplicial geometry for compositional data. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn (Eds.), *Compositional Data Analysis: from theory to practice*, Number 264 in Special Publications, pp. 145-160. The Geological Society, London, UK.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35* (3), 279-300.

Filzmoser, P. and K. Hron (2011). Robust statistical analysis. See Pawlowsky-Glahn and Buccianti (2011), pp. 59-72. 378 p.

Galton, F. (1879). The geometric mean, in vital and social statistics. *Proceedings of the Royal Society of London 29*, 365-366.

Martín-Fernández, J. A., J. Palarea, and R. Olea (2011). Dealing with zeros. See Pawlowsky-Glahn and Buccianti (2011), pp. 43-58. 378 p.

Mateu-Figueras, G. (2003). *Models de distribució sobre el símplex*. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2011). The principle of working on coordinates. See Pawlowsky-Glahn and Buccianti (2011), pp. 31-42. 378 p.

McAlister, D. (1879). The law of the geometric mean. *Proceedings of theRoyal Society of London 29*, 367-376.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate $\beta$-distribution and correlations among proportions. *Biometrika 49* (1-2), 65-82.
Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. See Thió-Henestrosa and Martín-Fernández (2003). CD-ROM.

Pawlowsky-Glahn, V. and A. Buccianti (Eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons. 378 p.

Pawlowsky-Glahn, V. and J. Egozcue (2011). Exploring compositional data with the coda-dendrogram. *Austrian Journal of Statistics 40* (1 & 2), 103-113.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA) 15* (5), 384-398.

Pawlowsky-Glahn, V. and J. J. Egozcue (2002). BLU estimators and compositional data. *Mathematical Geology 34* (3), 259-274.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London LX*, 489-502.

Sarmanov, O. V. and A. B. Vistelius (1959). On the correlation of percentage values. *Doklady of the Academy of Sciences of the USSR - Earth Sciences Section 126*, 22-25.

Thió-Henestrosa, S. and J. A. Martín-Fernández (Eds.) (2003). *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/. CD-ROM.

van den Boogaart, K., J. Egozcue, and V. Pawlowsky-Glahn (2010). Bayes linear spaces. *SORT - Statistics and Operations Research Transactions 34* (2), 201-222.