Mathematical Geology, Vol. 37, No. 7, October 2005 (© 2005)
DOI: 10.1007/s11004-005-7383-7

Compositional Data Analysis: Where Are We and Where Should We Be Heading?[1]

J. Aitchison[2] and J. J. Egozcue[3]

We take stock of the present position of compositional data analysis, of what has been achieved in the last 20 years, and then make suggestions as to what may be sensible avenues of future research. We take an uncompromisingly applied mathematical view, that the challenge of solving practical problems should motivate our theoretical research; and that any new theory should be thoroughly investigated to see if it may provide answers to previously abandoned practical considerations.

KEY WORDS: simplex geometry, Hilbert and Euclidean space, subcomposition, regression, sample space, stay-in-the-simplex.

INTRODUCTION

As stated in a previous version of the present paper (Aitchison, 2003), the statistical analysis of compositional data has gone through roughly four phases. The pre-1960 phase rode on the crest of the developmental wave of standard multivariate statistical analysis, an appropriate form of analysis for the investigation of problems with real sample spaces. Despite the obvious fact that a compositional vector—with components the proportions of some whole—is subject to a constant-sum constraint, and so is entirely different from the unconstrained vector of standard unconstrained multivariate statistical analysis, scientists and statisticians alike seemed almost to delight in applying all the intricacies of standard multivariate analysis, in particular correlation analysis, to compositional vectors. We know that Karl Pearson, in his definitive 1897 paper on spurious correlations, had pointed out the pitfalls of interpretation of such activity, but it was not until around 1960 that specific condemnation of such an approach emerged.

[1]Received 14 March 2004; accepted 28 February 2005.
[2]Department of Statistics, University of Glasgow, Scotland, UK; e-mail: john.aitchison@btinternet.com
[3]Dept. Matemática Aplicada III, Universitat Politècnica de Catalunya, Jordi Girona Salgado 1–3, Edicifi C2, E-08034 Barcelona, Spain; e-mail: juan.jose.egozcue@upc.edu

In the second phase, the primary critic of the application of standard multivariate analysis to compositional data was the geologist Chayes (1960), whose main criticism was in the interpretation of product–moment correlation between components of a geochemical composition, with negative bias the distorting factor from the viewpoint of any sensible interpretation. For this problem of negative bias, often referred to as the closure problem, Sarmanov and Vistelius (1959) supplemented the Chayes criticism in geological applications and Mosimann (1962) drew the attention of biologists to it. However, even conscious researchers, instead of working toward an appropriate methodology, adopted what can only be described as a pathological approach: distortion of standard multivariate techniques when applied to compositional data was the main goal of study.

The third phase was the realization by Aitchison in the 1980s that compositions provide information about relative, not absolute, values of components, that therefore every statement about a composition can be stated in terms of ratios of components (Aitchison, 1981, 1982, 1983, 1984). The facts that logratios are easier to handle mathematically than ratios and that a logratio transformation provides a one-to-one mapping onto a real space led to the advocacy of a methodology based on a variety of logratio transformations. These transformations allowed the use of standard unconstrained multivariate statistics applied to transformed data, with inferences translatable back into compositional statements.

The fourth phase arises from the realization that the internal simplicial operation of perturbation, the external operation of powering, and the simplicial metric, define a metric vector space (indeed a Hilbert space) (Billheimer, Guttorp, and Fagan, 1997, 2001; Pawlowsky-Glahn and Egozcue, 2001). So, many compositional problems can be investigated within this space with its specific algebraic–geometric structure. There has thus arisen a staying-in-the-simplex approach to the solution of many compositional problems (Mateu-Figueras, 2003; Pawlowsky-Glahn, 2003). This staying-in-the-simplex point of view proposes to represent compositions by their coordinates, 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.

The third phase, which mainly deals with (logratio) transformation of raw data, deserves special attention because these techniques have been very popular and successful over more than a century; from the Galton–McAlister introduction of such an idea in 1879 in their logarithmic transformation for positive data, through variance-stabilizing transformations for sound analysis of variance, to the general Box–Cox transformation (Box and Cox, 1964) and the implied transformations in generalized linear modelling. The logratio transformation principle was based on the fact that there is a one-to-one correspondence between compositional vectors and associated logratio vectors, so that any statement about compositions

can be reformulated in terms of logratios, and vice versa. The advantage of the transformation is that it removes the problem of a constrained sample space, the unit simplex, to one of an unconstrained space, multivariate real space, opening up all available standard multivariate techniques. The 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). The logratio methodology, however, drew fierce opposition from other disciplines, in particular from sections of the geological community. The reader who is interested in following the arguments that have arisen should examine the Letters to the Editor of Mathematical Geology over the period 1988 through 2002.

PRINCIPLES OF COMPOSITIONAL DATA ANALYSIS

Representation in the Simplex and Geometry

Two main principles of compositional data analysis are *scale invariance* and *subcompositional coherence*. Scale invariance merely reinforces the intuitive idea that a composition provides information only about relative values not about absolute values and, therefore, ratios of components are the relevant entities to study. This concept is easily formalized into a statement that all meaningful functions of a composition can be expressed in terms of a set of component ratios (Aitchison, 1997, 2002). Subcompositions of compositions are the analog of marginals or subvectors in unconstrained multivariate analysis (Aitchison, 1986, p. 33). Subcompositional coherence demands that two scientists, one using full compositions and the other using subcompositions of these full compositions, should make the same inference about relations within the common parts. Working with ratios, or equivalently logratios, involves not only scale invariance but automatically subcompositional coherence, since ratios within a subcomposition are equal to the corresponding ratios within the full composition. For details of these arguments associated with subcompositional coherence see Aitchison (1992b, 1994, 1997, 2002).

The *simplex of D parts*, $\mathcal{S}^D$, includes all positive real vectors adding up to a constant that for simplicity we take as the unit. Accordingly, absolute values of components in a composition are meaningless unless they are compared—by ratios—with other components. We use the notation $\mathcal{S}^D$, where the superscript is the number of parts of the composition. However, this subscript has been also used to indicate the dimension of the space, being $D - 1$.

Time has revealed the great importance of the basic operation of *perturbation* within the simplex (Aitchison, 1986, p. 42) in the analysis of compositional data.

Perturbation is computed by multiplying compositions component to component and, afterwards, dividing each component by the sum of all of them to attain unit sum. This last normalization is called *closure* and does not affect the ratios between components. Perturbation of two elements of the simplex of D parts, $\mathcal{S}^D$, $\mathbf{x}$ and $\mathbf{y}$, is often denoted by $\mathbf{x} \oplus \mathbf{y}$.

Perturbation has a *neutral element*. This is a composition with equal components; closure makes these components to be $1/D$ in a D-part simplex with unit closure constant. Any composition perturbed by this neutral element remains unaltered. The inverse operation of perturbation, denoted $\mathbf{y} \ominus \mathbf{x}$, is merely dividing components of a composition, $\mathbf{y}$, by the corresponding components of the other composition $\mathbf{x}$; closure reduces the result to an element of the simplex.

The underlying reason for the importance of perturbation is that it plays a role in the simplex precisely analogous to displacement or translation in real space; it is a mechanism for recording change. For example, if a D-part composition $\mathbf{x}$ changes through whatever process to a D-part composition $\mathbf{y}$, the change can be ascribed to a perturbation $\mathbf{p}$, with solution provided in terms of the inverse perturbation operator, $\mathbf{p} = \mathbf{y} \ominus \mathbf{x}$. Perturbation thus plays an important role not only in simple change as just described, but also in describing imprecision, in the definition and computation of residual compositions in compositional regression, and in other compositional fitting techniques. From the mathematical point of view, the perturbation operation defines an Abelian group on the simplex. This is the analog of translation in real spaces.

There is a second operation in the simplex, powering, the analog of scalar multiplication in real space. Powering a composition by a real constant consists of raising each component to the constant and then applying closure to the result. Notation for powering has changed several times but at this time we prefer to use $a \odot \mathbf{x}$, where a is a real constant and $\mathbf{x}$ a composition. Other alternative symbols have been used, e.g. $\otimes$ or $\diamond$, but $\odot$ seems to agree both with notation for perturbation and the familiar multiplication by scalars in real space. The operations $\oplus$ and $\odot$ define a $D-1$ dimensional vector or linear space structure on $\mathcal{S}^D$ (Pawlowsky-Glahn and Egozcue, 2002).

The structure can be extended to produce a metric vector space by the introduction of the *simplicial metric or distance* Δ_S defined by Aitchison (1983) as

$$\Delta_S(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{D} \left\{ \ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{y_i}{g(\mathbf{y})} \right\}^2 \right]^{1/2}$$

where $g(\cdot)$ denotes the geometric mean of the components of the enclosed vector. This distance has desirable properties, such as permutation and perturbation

invariance, a powering effect analogous to a scalar multiplication effect in real spaces and subcompositional dominance, which are relevant and indeed logically necessary for meaningful statistical analysis of compositional data (Aitchison, 1992a). This constitutes $\mathcal{S}^D$ as a metric space. A *norm*, $\|\mathbf{x}\|_S$, and *inner product*, $\langle \mathbf{x}, \mathbf{y} \rangle_S$, consistent with this metric, complete the Euclidean structure of the simplex. 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 (Billheimer, Guttorp, and Fagan, 1997, 2001; Egozcue and others, 2003; Pawlowsky-Glahn and Egozcue, 2001, 2002).

As for any vector space, generating vectors, bases, linear dependence, orthonormal bases, and subspaces play a fundamental role and this is equally true for the simplex metric vector space. In such concepts the counterpart of *linear combination* is a *power–perturbation combination*. This means that a composition can be represented by its coordinates with respect to a basis. For instance, if $\mathbf{b}_i$, $i = 1, 2, \ldots, D-1$, are independent compositions with respect to operations of perturbation and power, i.e., a basis of $\mathcal{S}^D$, any composition can be expressed as

$$\mathbf{x} = \left(x_1^* \odot \mathbf{b}_1\right) \oplus \left(x_2^* \odot \mathbf{b}_2\right) \oplus \ldots \oplus \left(x_{D-1}^* \odot \mathbf{b}_{D-1}\right),$$

where the asterisk denotes a coordinate.

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, $\mathbf{b}_1, \ldots, \mathbf{b}_{D-1}$, are $x_i^* = \langle \mathbf{x}, \mathbf{b}_i \rangle_S$, and they are called coordinates with respect to that basis. Coordinates are logratios and they were termed isometric logratios (ilr) since the corresponding transformation preserves the simplicial metric in $\mathcal{S}^D$ (Egozcue and others, 2003). The transformation that assigns the coordinates x_i^*, $i = 1, \ldots, D-1$, to the composition $\mathbf{x}$, allows us the computation of distances Δ_S, norms $\|\cdot\|_S$ and inner products $\langle \cdot, \cdot \rangle_S$ as ordinary Euclidean ones when using the coordinate vectors, e.g.

$$\Delta_S(\mathbf{x}, \mathbf{y}) = \sqrt{\left(x_1^* - y_1^*\right)^2 + \cdots + \left(x_{D-1}^* - y_{D-1}^*\right)^2},$$

where x_i^* and y_i^* are the coordinates of $\mathbf{x}$ and $\mathbf{y}$ with respect to the same orthonormal base. Within the ilr framework we can get different transformations corresponding to different orthonormal bases. Coordinates are by definition orthogonal logcontrasts (Aitchison, 1986, p. 85), involving ratios of compositional components in a more complicated way than simple logratios and so may pose more difficult problems in interpretation.

Later, we shall see that selection of adequate orthonormal bases plays a central role in a data-analytic sense in terms of the simplicial singular value decomposition of a compositional data set.

Clearly, in compositional processes, rates of change of compositions are important and here we review the basic ideas. Suppose that a composition depends on some continuous variable t such as time, length, or depth. Then, the rate of change of the composition with respect to t, or simplicial derivative, can be defined as the limit

$$D\mathbf{x}(t) = \lim_{dt \to 0} \frac{1}{dt} \odot [\mathbf{x}(t + dt) \ominus \mathbf{x}(t)] = \mathcal{C}\left[\exp\left(\frac{d}{dt} \ln \mathbf{x}(t)\right)\right] ,$$

where d/dt denotes ordinary differentiation with respect to t. This derivative has all the standard properties of derivatives. Similarly, simplicial integration with respect to t can be defined. These definitions are equivalent to representing the compositional function $\mathbf{x}(t)$ by its coordinates in some orthonormal basis of the simplex; performing the calculus (derivatives, integrals, etc.) using the coordinate representation; and, then, returning to the compositional representation. This equivalence is the principle of the staying-in-the-simplex way of dealing with compositional data. For further details of this algebraic–geometric structure of the simplex see Pawlowsky-Glahn and Egozcue (2001, 2002), Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn (2001), Aitchison (2002), Aitchison and others (2002), and Egozcue and others (2003).

Probability, Statistics, and Data Analysis

A first important step in any statistical problem is to determine a suitable sample space for available data. Frequently, real spaces are adequately adopted as sample spaces, thus providing both geometric background and sigma-field structure needed for defining events in a probabilistic framework. However, the nature of compositional data prevents the adoption of such an automatic sample space. Simplicial geometry has been shown to represent the main features of compositional data well and it is, therefore, the natural candidate to be chosen as the sample space of random compositional data. A necessary consequence of such a selection is the need to use probability models (distributions, densities) supported on the simplex and to ensure that events to which probability is relevant are also included in the space. Moreover, the use of the simplex as sample space invites us to use simplicial geometry when describing and modelling compositions.

A second step is to define a measure of central tendency. Inspired on Kullback–Leibler directed divergence concept, Aitchison (1997) defined the *center* of a random composition. Now, we can consider that definition as a consequence

of using simplicial geometry. It can be defined as the point, $\boldsymbol{\xi}$, in the simplex that minimizes the mean value of the squared simplicial distance, $\mathrm{E}[\Delta_S^2(\mathbf{X}, \boldsymbol{\xi})]$, i.e. $\mathrm{cen}[\mathbf{X}] = \boldsymbol{\xi}$ (Pawlowsky-Glahn and Egozcue, 2001, 2002). Therefore, the center plays the role of the expected value in real sample spaces. The center can be expressed as

$$\mathrm{cen}[\mathbf{X}] = \mathcal{C}[\exp(\mathrm{E}[\ln \mathbf{X}])] \,,$$

also called the geometric mean. A natural estimator of the center, $\hat{\boldsymbol{\xi}}$, is simply the closed vector of the geometric means of each component of the composition running over the sample. Consistent with this mean value there are a variety of equivalent measures of dispersion and covariance: the logratio covariance matrix (Aitchison, 1986, p. 76), the centred logratio covariance matrix (Aitchison, 1986, p. 79), and the variation matrix (Aitchison, 1986, p. 76). Importantly, these dispersion characteristics are consistent with the simplicial metric defined above and the estimator $\hat{\boldsymbol{\xi}}$ of the center is a best linear unbiased estimator from a stay-in-the-simplex point of view (Pawlowsky-Glahn, 2003; Pawlowsky-Glahn and Egozcue, 2002).

A further result in compositional data analysis based on simplicial geometry is the simplicial version of the singular value decomposition, on which much of multivariate statistical methodology is based. Consider a compositional data set, typically an $N \times D$ matrix, with n-th row composition $\mathbf{x}_n$. Any compositional data matrix can be decomposed in a power–perturbation form as follows

$$\mathbf{x}_n = \hat{\boldsymbol{\xi}} \oplus (u_{n1} p_1 \odot \mathbf{b}_1) \oplus (u_{n2} p_2 \odot \mathbf{b}_2) \oplus \cdots \oplus (u_{nR} p_R \odot \mathbf{b}_R) \,,$$

where $\hat{\boldsymbol{\xi}}$ is the estimate of the center of the data set; p_i, $i = 1, 2, \ldots, R$, are positive *singular values* in descending order of magnitude; $\mathbf{b}_i$, $i = 1, 2, \ldots, R$, are orthonormal compositions; R is a readily defined rank of the compositional data set; and the u's are power components specific to each $\mathbf{x}_n$. In practice, R is commonly $D - 1$, the full dimension of the simplex. In a way similar to that for real data sets, we may consider an approximation of order $r < R$ to the compositional data set given by

$$\mathbf{x}_n^{(r)} = \hat{\boldsymbol{\xi}} \oplus (u_{n1} p_1 \odot \mathbf{b}_1) \oplus (u_{n2} p_2 \odot \mathbf{b}_2) \oplus \cdots \oplus (u_{nr} p_r \odot \mathbf{b}_r), \;\; r < R \leq D - 1 \,.$$

Such an approximation retains a proportion $(p_1^2 + p_2^2 + \cdots + p_r^2)/(p_1^2 + p_2^2 + \cdots + p_R^2)$ of the total variability of the $N \times D$ compositional data matrix as measured by the trace of the estimated centered logratio covariance matrix or,

equivalently, in terms of the total mutual squared distances as

$$\frac{1}{N(N-1)} \sum_{m<n}^{D} \Delta_S^2(\mathbf{x}_n, \mathbf{x}_m).$$

We should remark that the singular value decomposition of a compositional data matrix is equivalent to identify orthogonal directions in the simplex such that the first one is the maximum variability direction of the sample; remaining variability has maximum variability in the second direction, and so on. Moreover, singular value decomposition in the simplex identifies a particular orthonormal basis in the simplex $\mathbf{b}_i$, $i = 1, 2, \ldots, R$, that should be completed up to $D - 1$ with other arbitrary orthonormal vectors if necessary. The coordinates, i.e., the ilr transform, of each composition $\mathbf{x}_n$ in the data matrix with respect to that orthonormal basis are also provided: they are $u_{n1}p_1, u_{n2}p_2, \ldots, u_{nR}p_R$, completed with null coordinates if the basis was completed up to $D - 1$ elements.

One of the most relevant consequences of expressing a composition by its coordinates with respect to some basis, preferably orthonormal, is that many standard multivariate problems that are stated in the simplex can be translated into coordinates and then formulated in the real space of those coordinates. A very important example is the comparison of the centers of two or more populations. Once all compositions have been expressed in coordinates, the problem reduces to a standard analysis of variance problem. These multivariate problems were initially discussed from the logratio transformation point of view (Aitchison and Ng, 2003; Buccianti, Nardi, and Potenza, 2003; Buccianti and Pawlowsky-Glahn, 2003; Pawlowsky-Glahn and Buccianti, 2002), but can now be thought of as a routine consequence of the staying-in-the-simplex methodology.

Beyond these algebraic–geometric applications to data analysis, adoption of the simplex as the sample space has required the study of distributions supported on the simplex. Among classical distributions, the Dirichlet distribution was initially the only one directly defined on the simplex. Various classes of parametric distributions on the simplex were consequently studied: the additive logistic normal (ALN) (Aitchison, 1986, p. 113), the multiplicative logistic normal (Aitchison, 1986, p. 130), the partitioned classes (Aitchison, 1986, p. 132), and the Dirichlet-embracing generalization ((Aitchison, 1985, 1986). A welcome addition is the multivariate logistic skew normal (Mateu-Figueras, 2003; Mateu-Figueras, Barceló-Vidal, and Pawlowsky-Glahn, 1998) based on the multivariate skew normal class introduced by Azzalini and Dalla Valle (1996) and further developed by Azzalini and Capitanio (1999). This allows for skewness in the logratio transformed data and promises more extensive study of methods which depend on distributional forms. For some uses of this distribution in compositional data

analysis see Aitchison and Bacon-Shone (1999) and Mateu-Figueras, Barceló-Vidal, and Pawlowsky-Glahn (1998).

Testing for distributional form and outlier detection has also been developed (Aitchison, Mateu-Figueras, and Ng, 2004; Barceló, Pawlowsky-Glahn, and Grunsky, 1996). It is also worth remembering that kernel density estimation is available for compositional data (Aitchison and Lauder, 1985).

The increased understanding of the algebraic–geometric structure of the underlying simplex sample space has opened up the possibility of a staying-in-the-simplex approach to compositional data analysis, as within a metric vector space, ideas such as minimum variance, unbiasedness, and least squares estimation, are available (Pawlowsky-Glahn, 2003; Pawlowsky-Glahn and Egozcue, 2002). It consists in using all geometric properties of the simplex as a Hilbert space. The consequences of this approach can be summarized as:

(a) Compositions have a double representation: the *raw* one, using vectors of parts, and the coordinate representation in some orthonormal basis. Log-ratios are not viewed as transformations of compositions but as coordinates or as *linear* combinations of them.
(b) The sample space of random compositions is the simplex and probability distributions of random compositions are supported on it.
(c) The simplicial metric (distance) between compositions demands consistent definitions of central tendency and dispersion measures.
(d) The simplicial measure (associated with the simplicial metric) is adopted to define probability densities. The usual Lebesgue measure associated with the Euclidean metric is no longer used as a reference. It only appears when dealing with coordinates, which are in a real space.

Let us illustrate item (d) with an example. Assume that $D = 3$ and that a composition $\mathbf{x} = [x_1, x_2, x_3]$ is represented by two orthonormal coordinates, e.g.,

$$\mathbf{x}^* = [x_1^*, x_2^*] \,, \quad \text{with} \quad x_1^* = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2} \,, \; x_2^* = \frac{1}{\sqrt{6}} \ln \frac{x_1 x_2}{x_3^2} \,.$$

If $\mathbf{x}^*$ is assumed to follow a normal distribution, $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$, then $\mathbf{x}$ follows an ALN distribution (Mateu-Figueras, 2003). If we want to represent the probability density of $\mathbf{x}$ on a ternary diagram using the simplicial measure as a reference, i.e., the Radon–Nykodym derivative of the probability measure with respect to the simplicial measure, we get the probability density

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{2\pi \, |\boldsymbol{\Sigma}^*|^{1/2}} \exp\left(-\frac{1}{2} \, \mathbf{x}^* \, (\boldsymbol{\Sigma}^*)^{-1} \, \mathbf{x}^{*t} \right) \,,$$

where the coordinate vector $\mathbf{x}^*$ is interpreted as a function of $\mathbf{x}$, the corresponding composition. Contours of this density are easily computed. We say that this kind of probability density corresponds to a normal distribution on $\mathcal{S}^D$ from a stay-in-the-simplex point of view. We realize that no Jacobian appears in this probability density; contrarily, if we take the usual Lebesgue measure on the simplex as reference measure we need to insert the Jacobian of the transformation from $\mathbf{x}$ into $\mathbf{x}^*$ in the expression of the density.

The experience of researchers in compositional data analysis has some lessons for workers with other forms of data. The importance of the identification of the principles such as scale invariance and subcompositional coherence, the clear definition of an appropriate sample space and recognition of the basic operations of change such as perturbation and power, have led us to meaningful systems of statistical inference. The same has been true of the analysis of directional data based on the special algebraic–geometric structure of the sphere. It is now being recognized that many, even most, standard multivariate data problems are concerned with positive (or nonnegative) vectors and that perhaps we should pay particular attention to the peculiar properties of the appropriate sample space, as discussed in Mateu-Figueras, Pawlowsky-Glahn, and Martín-Fernández (2002), Pawlowsky-Glahn and Egozcue (2002), Mateu-Figueras and Pawlowsky-Glahn (2003), Pawlowsky-Glahn (2003), Pawlowsky-Glahn, Egozcue and Burger (2003), and Tolosana-Delgado, Pawlowsky-Glahn and Mateu-Figueras (2003).

QUESTIONS, MODELS, AND APPLICATIONS

Limitations in the Interpretability of Compositional Data

There is a tendency in some compositional data analysts to expect too much in their inferences from compositional data. For these, the following situation may show the nature of the limitations of compositional data. We have a planter consisting of water, soil, and seed. One evening a sample was taken and its (water, soil, seed) composition was determined as $\mathbf{x}_1 = [3/6, 2/6, 1/6]$. In the morning again we analyzed a sample, finding $\mathbf{x}_2 = [6/9, 2/9, 1/9]$. We measured the change as the perturbation

$$\mathbf{p} = \mathbf{x}_2 \ominus \mathbf{x}_1 = \mathcal{C}\left[\frac{(3/6)}{(6/9)}, \frac{(2/6)}{(2/9)}, \frac{(1/6)}{(1/9)}\right] = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right].$$

Now we can picture two simple scenarios which could describe this change. Suppose that the planter last evening actually contained [18, 12, 6] kg of (water, soil, seed), corresponding to the evening composition [3/6, 2/6, 1/6], and it rained during the night increasing the water content only, so that the morning content

was [36, 12, 6] kg, corresponding to the morning composition [6/9, 2/9, 1/9]. Although this rain-only explanation may be true, the change could equally be explained by a wind-only scenario, in which the overnight wind had swept away soil and seed resulting in content of [18, 6, 3] kg and the same morning composition [6/9, 2/9, 1/9]. Even more complicated scenarios will produce a similar change. The point here is that compositions provide information only about the relative magnitudes of the compositional components and so interpretations involving absolute values as in the above example cannot be justified. Only if there is evidence external to the compositional information would such inferences be justified. For example, if it is known that strong winds and no rain occurred that night, the wind-only scenario would be justified. A consequence of this example is that we must learn to phrase our inferences from compositional data in terms which are meaningful and we have seen that the meaningful operations are perturbation and powering.

This example may also illustrate the convenience of asking ourselves *is my problem only compositional?* Imagine that preservation of mass of seed is the goal of our analysis. If we only look to the compositional results we arrive at the (wrong) conclusion that wind and rain scenarios cause similar loss of seed. However, the mass data show that in the rain scenario no loss of seed occurred. Our particular goal is related to mass and then the problem is not only compositional. It requires the use of additional data such as masses in combination with compositions.

Compositional Processes

Most scientists are interested in the nature of the process which has led to the data they observe. For example, geological language contains many terms to describe a whole variety of envisaged processes, such as denudation, diagenesis, erosion, gravity transport, metasomatism, metamorphism, mixing, orogenesis, polymetamorphism, sedimentation, transportation, weathering. Unfortunately, the scientist is seldom in the position of observing a closed system where fundamental principles such as conservation of mass and energy apply. Commonly the only data available take the form of compositional data providing information only on relative magnitudes of the constituents of the specimens, e.g., percentages by weight or chemical concentrations. Thus, there is a need to extend compositional data analysis to provide satisfactory models to describe such processes.

The calculus available in the simplex—perturbation, powering, derivatives, and integrals—allows us to propose sound models for processes influencing composition. The simplest one is the compositional straight-line, parametrized by some external variable like time, space, or depth. For instance a t-dependent composition may be modelled as $\mathbf{x}(t) = \mathbf{x}_0 \oplus (t \odot \mathbf{p})$, where $\mathbf{x}_0$ is a point in the line, eventually the initial condition, and $\mathbf{p}$ the direction of the line. This simple process is the general solution of the compositional differential equation $D\mathbf{x}(t) = \mathbf{p}$. Lines

in the simplex can be interpreted as compositional evolution of an exponential mass growth or decay. The simplicity of lines in the simplex promotes attempts to discover simple *natural laws* from compositional observations. The tools for such discovery are again either principal logcontrast analysis or, equivalently, singular value decomposition. For details of such discoveries through principal logcontrast analysis see Aitchison and Thomas (1998), Aitchison (1999), Buccianti and others (1999), Egozcue and others (2003), Thomas and Aitchison (2003), von Eynatten, Barceló-Vidal, and Pawlowsky-Glahn (2003) and through biplot analysis see Aitchison and Greenacre (2002). However, further work is required in this field because, in fact, differential models in the simplex may provide models exceeding the simplicity of the line such as oscillations or periodicities.

Compositional Regression

Compositional regression, where the composition is the regressand and we seek an explanation of its variability in terms of factors and/or a concomitant variable, has been extensively discussed and illustrated in Aitchison (1986, 7.6–7.9) and need not be further discussed here. Such linear modelling within the logratio transformation methodology is simple and can rely on standard multivariate techniques. The expression of compositional regression by the staying-in-the-simplex approach is by way of power–perturbation combinations. A composition $\mathbf{y}$ depends on compositional concomitants $\mathbf{x}_1, \mathbf{x}_2, \ldots$ as

$$\mathbf{y} = \mathbf{a} \oplus (t_1 \odot \mathbf{x_1}) \oplus (t_2 \odot \mathbf{x_2}) \oplus \cdots \oplus \mathbf{p} \,,$$

where the composition $\mathbf{a}$ is the analog of the *intercept* in ordinary regression, the real constants $t_1, t_2, \ldots$ are the analogs of the regression coefficients, and $\mathbf{p}$ is the perturbation error. Clearly, interpretation here is dependent on a sound mathematical appreciation of the algebraic–geometric structure of the simplex.

With actual compositional data, the regression either in logratio terms or in staying-in-the-simplex mode is easily accomplished. The important feature here is the possibility of alternative approaches to interpretation. For further details and an application see Aitchison and Thomas (1998) and for further developments see Aitchison and Barceló-Vidal (2002) and Daunis-i-Estadella, Egozcue and Pawlowsky-Glahn (2002).

Binary Logistic Discrimination

The classification-diagnostic problem illustrates how this technique may be developed. For two classes of compositions, $(\theta = 0, \theta = 1)$, $\mathbf{x} = [x_1, x_2, \ldots, x_D]$,

a useful model is the binary logistic model, using a logcontrast

$$\text{lc}(\mathbf{a}, \mathbf{x}) = a_0 + a_1 \ln x_1 + a_2 \ln x_2 + \cdots + a_D \ln x_D, \quad \sum_{i=1}^{D} a_i = 0$$

as the regressor. More specifically,

$$\text{P}[\theta = k|\mathbf{x}] = \frac{\exp[\text{lc}(\mathbf{a}, \mathbf{x})]}{1 + \exp[\text{lc}(\mathbf{a}, \mathbf{x})]}, \quad k = 0, 1 .$$

Maximum likelihood estimation of the parameter **a** is straightforward. The beauty of this model is that the adequacy of a subcomposition (say the parts $1, \cdots, C$) as a substitute for the complete composition in the process of classification can readily be tested since this hypothesis can be expressed as $a_{C+1} = a_{C+2} = \cdots = a_D = 0$. Thus the whole lattice of subcompositional hypotheses can be investigated and any adequate subcomposition identified (Aitchison, 1986, 12.6, 12.7; Thomas and Aitchison, 1998, 2003).

Extensions to more classes of compositions are straightforward. An example of this is *sequential discrimination*. Even with more than two classes, the above binary logistic regression approach may be possible, even sensible. A geological example of related techniques can be found in Tolosana-Delgado and others (2002).

Convex Linear Mixing Processes

A popular way of studying compositional data, such as sedimentology and environmental pollution studies, is in terms of *convex linear mixing* processes. Such an approach is based on some such assumption as conservation of mass. There is, of course, no way that compositional data can be used to support such a mass conservation hypothesis since compositions carry no information about mass. Compositions can, however, be analyzed within models which assume conservation of mass. All these models assume that there are source compositions, say $\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_C$, from which a generic observed composition arises as a convex linear combination

$$\mathbf{x} = a_1\mathbf{z}_1 + a_2\mathbf{z}_2 + \cdots + a_C\mathbf{z}_C ,$$

where $\mathbf{a} = [a_1, a_2, \ldots, a_C] \in \mathcal{S}^C$ is the vector of mixing proportions. The form of modelling obviously depends on the extent of the information about the number of sources and the source compositions. In some cases, neither the number of sources nor their compositions are known, the so-called *endmember* problem as discussed,

for example, by Renner (2003) and Weltje (1997). At the opposite extreme, the problem may be to test a hypothesis that the sources are specified compositions $\mathbf{z}_1$, $\mathbf{z}_2, \ldots \mathbf{z}_C$. Many intermediate situations can be visualized: an example is the pollution problem analyzed by Aitchison and Bacon-Shone (1999), where there are not only samples from the target set but also sampled compositions from the sources. Note that the basic operation here is an additive one, so that all the nice distributional properties associated with perturbation and powering are not available, i.e., mixing or convex linear combination is not a linear operation in the simplex. For example, given that $\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_C$ are independently distributed as additive logistic normal and that $\mathbf{a}$ is a constant vector or has some given logistic normal distribution, no explicit form for the distribution of the convex linear mixture can be found. It was only by the determination of good approximations to the distribution that Aitchison and Bacon-Shone (1999) resolved their pollution problem.

The additive nature of such modelling does not mean that basic principles of compositional data analysis should be neglected in these mixing problems. In solutions of the endmember problem there has been a tendency to avoid the simplicial metric and to revert to Euclidean distance and classical least squares in estimating mixture vectors. This is certainly not necessary and the more appropriate simplicial metric may be used. For example, an approach to the so-called endmember problem where a set of say C endmember compositions $\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_C$ is sought such that each composition $\mathbf{x}_n$, $n = 1, \ldots, N$, of the data set can be expressed as a convex linear combination of $\mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_C$. In order to find out suitable values of $\mathbf{a}$ the optimality criterion used may be expressed in terms of simplicial distances as $\Delta_S(\mathbf{x}_n, \mathbf{x}(\mathbf{a}))$, $n = 1, 2, \ldots, N$, where $\mathbf{x}(\mathbf{a})$ is a convex linear combination of the C endmember compositions and $\mathbf{a}$ is a vector containing the C unknown coefficients. See Aitchison and Barceló-Vidal (2002) for further details and an example of a method of comparing the adequacy of differential perturbation and convex linear mixing processes, or Tolosana-Delgado and others (2005) for a possible alternative approach based on simplicial factor analysis. In the computation for such analysis a basic algorithm is obviously required for the maximization or minimization of a function on the simplex and we now have efficient search algorithms based on perturbation techniques.

Graphical Aids

Visual representation of compositional data and their features requires special *graphical aids*. Available techniques may be summarized in some few categories as follows.

Harker and Related Diagrams

It is now over four decades since Felix Chayes warned geologists of the dangers of attempting to interpret Harker and similar diagrams where one component

of a composition is plotted against another. Yet a recent search of the web under *Harker diagram* produced some 60 sites, many of them instructing students in the use of such a *graphical tool*. The only legitimate use of such diagrams is in terms of the ratios, that is in terms of the rays from the origin to the data points. In our view, Harker diagrams are best condemned as misleading and best left out of any attempts to interpret compositional variability.

Ternary Diagrams

Like Harker diagrams these should be treated with caution. For example, in the past there has been substantial discussion on the nature of data sets with apparent curvature within a ternary diagram (Butler, 1979). Are these trends or not? With our knowledge of the algebraic–geometric structure of the simplex we now know that constant logcontrast curves are indeed the straight lines of the simplex and so any interpretation of curvature within the ternary diagram should be treated with substantial care. See Aitchison and Thomas (1998) and Buccianti and others (1999) for examples where such curvature can be interpreted as a trend or compositional process. However, ternary diagrams, complemented with centering and scaling techniques (Martín-Fernández and others, 1999; von Eynatten, Pawlowsky-Glahn and Egozcue, 2002) are one of the most important and practical tools to represent compositional data.

Ratio and Logratio Scatter-Plots

If scatter-plots are to be used in interpreting compositional data then because of the necessity to meet the demands of the principle of scale invariance they should involve ratios or logratios. A good example of how such diagrams can be used for exposition is to be found in the discriminatory example in Thomas and Aitchison (1998). But care should be taken when visualizing distances because they may be different from the simplicial ones, specially when a common denominator is used.

Compositional Biplots

The development of biplot techniques for compositional data is a substantial advance in the study of compositional data sets. The biplot (Gabriel, 1971, 1981) is a well-established graphical aid in other branches of statistical analysis. Its adaptation for compositional data is simple and can prove a useful exploratory and expository tool. For a compositional data set the biplot is based on a singular value decomposition of the doubly centered logratio matrix. For details of biplot construction see Aitchison (1990, 1997, 2002) and Aitchison and Greenacre

(2002). Such biplots, consisting of vertices, rays, links, and case markers, allow an overall view of compositional covariance structure, subcompositional analysis, the relationship of individual compositions to parts, and provide useful interpretations of near-coincident vertices, collinear vertices, and orthogonal links. There are obvious extensions of biplot methodology to bicompositions and to conditional biplots.

Coordinate Scatter-Plots

They consist in representing coordinates of compositions with respect to some orthonormal basis in a 1D, 2D, or 3D plot. They are a special class of logratio scatter-plots when the selected logratios are orthonormal. Ordinary Euclidean distance on the plot is equivalent to simplicial distance so overcoming scale problems in other kind of logratio scatter-plots (Egozcue and others, 2003). It should be pointed out, however, this gain in visualizing distance may in practice be offset by the fact that the coordinates are necessarily orthogonal logcontrasts and so may be substantially more complicated to interpret than simple logratios.

Problems of Zero Components

Neither logratio analysis nor the stay-in-the-simplex approach can deal with compositional data for which some component is recorded as a zero. Often this occurs because the recording device was unable to detect the proportion of such a part. In these cases replacement of the zero value by a small nonzero value may overcome the problem. The replacement method (Aitchison, 1986, p. 266) of rounded or trace zeros is not subcompositionally coherent and should now be replaced by the method arrived at independently by Fry, Fry, and McLaren (2000) and Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000), which preserve the ratios of nonzero components. Such nonparametric replacement procedures still appear to be the most viable methods available provided sensitivity analysis over a sensible range of replacement values is used as a check. For a parametric approach, see Martin-Fernández, Paladea-Albadalejo and Gómez-García (2003).

One of the tantalizing remaining problems in compositional data analysis lies in how to deal with data sets in which there are components which are *essential or structural zeros*. By an essential or structural zero we mean a component which is truly zero, not something recorded as zero simply because the experimental design or the measuring instrument has not been sufficiently sensitive to detect a trace of the part. Such essential zeros occur in many compositional situations; typical examples can be found in paleontological or palynological taxa. Devices such as nonzero replacement and amalgamation are almost invariably ad hoc and unsuccessful. An alternative approach through ranking of components is given by

Bacon-Shone (1992). For some essential or structural zeros careful consideration of the questions being asked can sometimes remove the problem; see for example the predator–prey example in Aitchison (1986, 11.7).

Research is under way to attempt to construct two-stage models for the treatment of essential or structural zeros. In such modelling it seems sensible to build up a model in two stages, the first determining where the zeros will occur and the second how the unit available is distributed among the nonzero parts. Two reports on this promising line of research were presented in CODAWORK-03 workshop by Aitchison and Kay (2003) and Bacon-Shone (2003).

Reduction of Dimensionality

The number of parts of large compositional vectors should be frequently reduced both to remove undesired information and to get interpretable results. *Singular value decomposition*, as presented in a previous section, is the main tool for such a reduction and is parallel to principal component analysis for real multivariate data. Data are projected on the first principal directions and the remaining components are removed or considered as residuals. This projection is an orthogonal one in the simplicial sense. The biplot is the result of such an approach when two principal components are chosen to represent the data. For an example see Tolosana-Delgado and others (2005).

Principal component analysis projects data into a subspace of a given dimension that is chosen to explain the most of the variance of data. However, there are alternative criteria to obtain a reduction of the dimension. The most standard one is that of *subcompositional analysis* which causes a drastic dimensionality reduction in a well-established framework (Aitchison, 1986, p. 33–73).

Another reduction technique is the partition analysis as described in Aitchison (1986, p. 38). A compositional vector is partitioned into groups of parts. The components within a group are amalgamated and these amalgamated parts constitute a new compositional vector of reduced dimension, with as many parts as groups were obtained in the partition. However, amalgamation of components in a compositional vector is not a linear operation from the simplicial point of view and such techniques are under discussion. Most arguments arose in the workshop on compositional data analysis, CODAWORK-03, held in Girona (Spain) 2003, where several contributors used amalgamation to reduce dimensionality of their data (Thió-Henestrosa and Martín-Fernández, 2003). An alternative to amalgamation of parts is presented in a paper in this issue (Egozcue and Pawlowsky-Glahn, 2005).

Almost all issues in standard multivariate statistical analysis may have their compositional version. While their use does not involve new computational problems because logratio analysis transform them into a standard one, interpretation

of results may introduce new challenging difficulties that should be discussed in the framework of their respective applications.

A VIEW OF THE FUTURE OF COMPOSITIONAL DATA ANALYSIS

We think that the interesting future of compositional data analysis will lie in statisticians searching for real applied problems in as many disciplines as possible. A recent search of the web under *compositional data* located over 3000 sites varying over a wide variety of disciplines, so there are plenty of challenges in this direction. Equally important is that applied workers in these disciplines should search out statisticians and present them with the challenge of answering their compositional questions. Tchebycheff, in his Theory of Maps has the fundamental idea: *Real progress is made when theory and the needs of application go hand in hand*.

A substantial problem for those of us who have tried to promote understanding of the special features of compositional data analysis has been the inertia of statistical practice established along more than a century. These resistances range from those who think that the simplex is nothing more than a subset of real space and *there isn't a special problem*, through those who insist that the new methodology should be doing little more than corroborating views already obtained and firmly held from previous analysis, to some who have used averages, raw correlations, and ordinary Euclidean distances for compositional data analysis and now realize that their work is being attacked.

These circumstances point out that a further effort should be made to explain the soundness of models based on the structure of the simplex and giving tools to interpret obtained results. Difficulties coming from the fact that representation of compositional data is done by coordinates, mainly logratios, and the interpretation remains in the original compositions, should be overcome and this can only be done by a rigorous and continued effort in applications that presumably range across all fields of science and technology.

REFERENCES

Aitchison, J., 1981, A new approach to null correlations of proportions: Math. Geol., v. 13, no. 2, p. 175–189.

Aitchison, J., 1982, The statistical analysis of compositional data (with discussion): J. R. Stat. Soc., Ser. B (Stat. Methodol.), v. 44, no. 2, p. 139–177.

Aitchison, J., 1983, Principal component analysis of compositional data: Biometrika, v. 70, no. 1, p. 57–65.

Aitchison, J., 1984, The statistical analysis of geochemical compositions: Math. Geol., v. 16, no. 6, p. 531–564.

Aitchison, J., 1985, A general class of distributions on the simplex: J. R. Stat. Soc., Ser. B (Stat. Methodol.), v. 47, no. 1, p. 136–146.

Aitchison, J., 1986, The statistical analysis of compositional data. Monographs on statistics and applied Probability: Chapman & Hall, London (Reprinted in 2003 with additional material by Press Blackburn), 416 p.

Aitchison, J., 1990, Relative variation diagrams for describing patterns of compositional variability: Math. Geol., v. 22, no. 4, p. 487–511.

Aitchison, J., 1992a, On criteria for measures of compositional difference: Math. Geol., v. 24, no. 4, p. 365–379.

Aitchison, J., 1992b, The triangle in statistics, *in* Mardia, K., ed., The art of statistical science. A tribute to G. S. Watson: Wiley, New York, p. 89–104.

Aitchison, J., 1994, Principles of compositional data analysis, *in* Anderson, T. W., Olkin, I., and Fang, K., eds., Multivariate analysis and its applications: Institute of Mathematical Statistics, Hayward, CA, p. 73–81.

Aitchison, J., 1997, The one-hour course in compositional data analysis or compositional data analysis is simple, *in* Pawlowsky-Glahn, V., ed., Proceedings of IAMG'97—The third annual conference of the International Association for Mathematical Geology, Vol. I, II and addendum: International Center for Numerical Methods in Engineering (CIMNE), Barcelona, Spain, p. 3–35.

Aitchison, J., 1999, Logratios and natural laws in compositional data analysis: Math. Geol., v. 131, no. 5, p. 563–580.

Aitchison, J., 2002, Simplicial inference, *in* Viana, M. A. G., and Richards, D. S. P., eds., Algebraic methods in statistics and probability, v. 287, Contemporary mathematics series: American Mathematical Society, Providence, RI, p. 1–22.

Aitchison, J., 2003, Compositional data analysis: Where are we and where should we be heading? See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Aitchison, J., and Bacon-Shone, J., 1999, Convex linear combination of compositions: Biometrika, v. 86, no. 2, p. 351–364.

Aitchison, J., and Barceló-Vidal, C., 2002, Compositional processes: A statistical search for understanding: See Bayer, Burger, and Skala (2002, p. 381–386).

Aitchison, J., Barceló-Vidal, C., Egozcue, J. J., and Pawlowsky-Glahn, V., 2002, A concise guide for the algebraic–geometric structure of the simplex, the sample space for compositional data analysis. See Bayer, Burger, and Skala (2002, p. 387–392).

Aitchison, J., and Greenacre, M., 2002, Biplots for compositional data: J. R. Stat. Soc., Ser. C (Appl. Stat.), v. 51, no. 4, p. 375–392.

Aitchison, J., and Kay, J., 2003, Possible solution of some essential zero problems in compositional data analysis. See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Aitchison, J., and Lauder, I. J., 1985, Kernel density estimation for compositional data: J. R. Stat. Soc., Ser. C (Appl. Stat.), v. 34, no. 2, p. 129–137.

Aitchison, J., Mateu-Figueras, G., and Ng, K. W., 2004, Characterization of distributional forms for compositional data and associated distributional tests: Math. Geol., v. 35, no. 6, p. 667–680.

Aitchison, J., and Ng, K. W., 2003, Compositional hypotheses of subcompositional stability and specific perturbation change and their testing. See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Aitchison, J., and Shen, S. M., 1980, Logistic-normal distributions. Some properties and uses: Biometrika, v. 67, no. 2, p. 261–272.

Aitchison, J., and Thomas, C. W., 1998, Differential perturbation processes: A tool for the study of compositional processes. See Buccianti, Nardi, and Potenza (1998, p. 499–504).

Azzalini, A., and Capitanio, A., 1999, Statistical applications of the multivariate skew-normal distribution: J. R. Stat. Soc., Ser. B (Stat. Methodol.) v. 61, no. 3, p. 579–602.

Azzalini, A., and Dalla Valle, A., 1996, The multivariate skew-normal distribution: Biometrika, v. 83, no. 4, p. 715–726.

Bacon-Shone, J., 1992, Ranking methods for compositional data: Appl. Stat., v. 41, no. 3, p. 533–537.

Bacon-Shone, J., 2003, Modelling structural zeros in compositional data. See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Barceló, C., Pawlowsky-Glahn, V., and Grunsky, E., 1996, Some aspects of transformations of compositional data and the identification of outliers: Math. Geol., v. 28, no. 4, p. 501–518.

Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V., 2001, Mathematical foundations of compositional data analysis, *in* Ross, G., ed., Proceedings of IAMG'01—The sixth annual conference of the International Association for Mathematical Geology, CO-ROM, 20 p.

Bayer, U., Burger, H., and Skala, W., eds., 2002, Proceedings of IAMG'02—The eighth annual conference of the International Association for Mathematical Geology, Terra Nostra, no. 3,

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

Billheimer, D., Guttorp, P., and Fagan, W., 2001, Statistical interpretation of species composition: J. Am. Stat. Assoc., v. 96, no. 456, p. 1205–1214.

Box, G. E. P., and Cox, D. R., 1964, The analysis of transformations: J. R. Stat. Soc., Ser. B (Stat. Methodol.), v. 26, no. 2, p. 211–252.

Buccianti, A., Nardi, G., and Potenza, R., eds., 1998, Proceedings of IAMG'98—The fourth annual conference of the International Association for Mathematical Geology, Vol. I and II: De Frede Editore, Napoli, 969 p.

Buccianti, A., and Pawlowsky-Glahn, V., 2003, Random variables and geochemical processes: A way to describe natural variability: *in* Ottonello, G., and Serva, L., Geochemical baselines of Italy, Chapter 4: Pacini Editore, Genova, Italy, 294 p.

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. See Lippard, Næss, and Sinding-Larsen (1999, p. 139–144).

Buccianti, A., Vaselli, O., and Nisi, B., 2003, New insights on river water chemistry by using non-centred simplicial principal component analysis: A case study. See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Butler, J. C., 1979, The effects of closure on the moments of a distribution: Math. Geol., v. 11, no. 1, p. 75–84.

Chayes, F., 1960, On correlation between variables of constant sum: J. Geophys. Res., v. 65, no. 12, p. 4185–4193.

Daunis-i-Estadella, J., Egozcue, J. J., and Pawlowsky-Glahn, V., 2002, Least squares regression in the simplex. See Bayer, Burger, and Skala (2002, p. 411–416).

Egozcue, J. J., and Pawlowsky-Glahn, V., 2005, Groups of parts and their balances in compositional data analysis. Math. Geol., v. 37, no. 7, p. 795–828.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C., 2003, Isometric logratio transformations for compositional data analysis: Math. Geol., v. 35, no. 3, p. 279–300.

Fry, J. M., Fry, T. R. L., and McLaren, K. R., 2000, Compositional data analysis and zeros in micro data: Appl. Econ., v. 32, no. 8, p. 953–959.

Gabriel, K. R., 1971, The biplot—graphic display of matrices with application to principal component analysis: Biometrika, v. 58, no. 3, p. 453–467.

Gabriel, K. R., 1981, Biplot display of multivariate matrices for inspection of data and diagnosis, *in* Barnett, V., ed., Interpreting multivariate data: Wiley, New York, p. 147–173.

Galton, F., 1879, The geometric mean, in vital and social statistics: Proc. R. Soc. Lond., v. 29, p. 365–366.

Lippard, S. J., Næss, A., and Sinding-Larsen, R., eds., 1999, Proceedings of IAMG'99—The fifth annual conference of the International Association for Mathematical Geology, Vol. I and II: Tapir, Trondheim, Norway, 784 p.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2000, Zero replacement in compositional data sets, *in* Kiers, H., Rasson, J., Groenen, P., and Shader, M., eds., Studies in classification, data analysis, and knowledge organization: Springer-Verlag, Berlin, p. 155–160.

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. See Lippard, Næss, and Sinding-Larsen (1999, p. 211–216).

Martin-Fernández, J. A., Paladea-Albadalejo, J., and Gómez-García, J., 2003, Markov chain Monte Carlo method applied to rounding zeros of compositional data: First approach. See Thió-Henestrosa and Martín-Fernández (2003) (electronic publication).

Mateu-Figueras, G., 2003, Models de distribució sobre el símplex: PhD Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

Mateu-Figueras, G., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1998, Modeling compositional data with multivariate skew-normal distributions. See Buccianti, Nardi, and Potenza (1998, p. 532–537).

Mateu-Figueras, G., and Pawlowsky-Glahn, V., 2003, Una alternativa a la distribución lognormal. See Saralegui and Ripoll (2003) (electronic publication).

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Martín-Fernández, J. A., 2002, Normal in $\Re^+$ vs. lognormal in $\Re$. See Bayer, Burger, and Skala (2002, p. 305–310).

McAlister, D., 1879, The law of the geometric mean: Proc. R. Soc. Lond., v. 29, p. 367–376.

Mosimann, J. E., 1962, On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions: Biometrika, v. 49, nos. 1–2, p. 65–82.

Pawlowsky-Glahn, V., 2003, Statistical modelling on coordinates. See (Thió-Henestrosa and Martín-Fernández, 2003) (electronic publication).

Pawlowsky-Glahn, V., and Buccianti, A., 2002, Visualization and modeling of subpopulations of compositional data: Statistical methods illustrated by means of geochemical data from fumarolic fluids: Int. J. Earth Sci. (Geol. Rundschau), v. 91, no. 2, p. 357–368.

Pawlowsky-Glahn, V., and Egozcue, J. J., 2001, Geometric approach to statistical analysis on the simplex: Stochastic Environ. Res. Risk Assess. (SERRA), v. 15, no. 5, p. 384–398.

Pawlowsky-Glahn, V., and Egozcue, J. J., 2002, BLU estimators and compositional data: Math. Geol., v. 34, no. 3, p. 259–274.

Pawlowsky-Glahn, V., Egozcue, J. J., and Burger, H., 2003, An alternative model for the statistical analysis of bivariate positive measurements, *in* Cubitt, J., ed., Proceedings of IAMG'03—The ninth annual conference of the International Association for Mathematical Geology, CD-ROM: University of Portsmouth, Portsmouth, UK.

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: Proc. R. Soc. Lond., v. LX, p. 489–502.

Renner, R. M., 1993, The resolution of a compositional data set into mixtures of fixed source components: J. R. Stat. Soc., Ser. C (Appl. Stat.), v. 42, no. 4, p. 615–631.

Saralegui, J., and Ripoll, E., eds., 2003, Actas del XXVII Congreso Nacional de la Sociedad de Estadística e Investigación Operativa (SEIO), CD-ROM: Sociedad de Estadística e Investigación Operativa, Lleida (Spain).

Sarmanov, O. V., and Vistelius, A. B., 1959, On the correlation of percentage values: Dokl. Akad. Nauk. SSSR, v. 126, p. 22–25.

Thió-Henestrosa, S., and Martín-Fernández, J. A., eds., 2003, Compositional Data Analysis Workshop—CoDaWork'03, Proceedings: Universitat de Girona, CD-ROM, ISBN 84-8458-111-X, available at http://ima.udg.es/Activitats/CoDaWork03/.

Thomas, C. W., and Aitchison, J., 1998, The use of logratios in subcompositional analysis and geochemical discrimination of metamorphosed limestones from the northeast and central scottish highlands. See Buccianti, Nardi, and Potenza (1998, p. 549–554).

Thomas, C. W., and Aitchison, J., 2003, Exploration of geological variability and possible processes through the use of compositional data analysis: An example using Scottish metamorphosed limestones. See Buccianti, Nardi, and Potenza (1998) (electronic publication).

Tolosana-Delgado, R., Otero, N., Pawlowsky-Glahn, V., and Soler, A., 2005, Extracting latent factor subcompositions from hydrochemical conpositions. Math. Geol., v. 37, no. 7, p. 681–702.

Tolosana-Delgado, R., Palomera-Román, R., Gimeno-Torrente, D., Pawlowsky-Glahn, V., and Thió-Henestrosa, S., 2002, A first approach to the classification of basalts using trace elements. See Bayer, Burger, and Skala (2002, p. 435–440).

Tolosana-Delgado, R., Pawlowsky-Glahn, V., and Mateu-Figueras, G., 2003, Krigeado de variables positivas. Un modelo alternativo. See Bayer, Burger, and Skala (2002) (electronic publication).

von Eynatten, H., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2003, Modelling compositional change: The example of chemical weathering of granitoid rocks: Math. Geol., v. 35, no. 3, p. 231–251.

von Eynatten, H., Pawlowsky-Glahn, V., and Egozcue, J. J., 2002, Understanding perturbation on the simplex: A simple method to better visualize and interpret compositional data in ternary diagrams: Math. Geol., v. 34, no. 3, p. 249–257.

Weltje, J. G., 1997, End-member modeling of compositional data: Numerical–statistical algorithms for solving the explicit mixing problem: Math. Geol., v. 29, no. 4, p. 503–549.