1

A short history of compositional data analysis

John Bacon-Shone
Social Sciences Research Centre, The University of Hong Kong, Hong Kong

1.1 Introduction

Compositional data are data where the elements of the composition are non-negative and sum to unity. While the data can be generated directly (e.g. probabilities), they often arise from non-negative data (such as counts, area, volume, weights, expenditures) that have been scaled by the total of the components. Geometrically, compositional data with D components has a sample space of the regular unit D-simplex, $\mathcal{S}^D$. The key question is whether standard multivariate analysis, which assumes that the sample space is $\mathbb{R}^D$, is appropriate for data from this restricted sample space and if not, what is the appropriate analysis? Ironically, most multivariate data are non-negative and hence already have a sample space with a restriction to $\mathbb{R}^D_+$. This chapter tries to summarize more than a century of progress towards answering this question and draws heavily on the review paper by Aitchison and Egozcue (2005).

1.2 Spurious correlation

The starting point for compositional data analysis is arguably the paper of Pearson (1897), which first identified the problem of '*spurious correlation*' between ratios of variables. It is easy to show that if X, Y and Z are uncorrelated, then X/Z and Y/Z will not be uncorrelated. Pearson then looked at how to adjust the correlations to take into account the '*spurious*

Compositional Data Analysis: Theory and Applications, First Edition. Edited by Vera Pawlowsky-Glahn and Antonella Buccianti.

correlation' caused by the scaling. However, this ignores the implicit constraint that scaling only makes sense if the scaling variable is either strictly positive or strictly negative. In short, this approach ignores the range of the data and does not assist in understanding the process by which the data are generated. Tanner (1949) made the essential point that a log transform of the data may avoid the problem and that checking whether the original or log transformed data follow a Normal distribution may provide some guidance as to whether a transform is needed.

Chayes (1960) later made the explicit connection between Pearson's work and compositional data and showed that some of the correlations between components of the composition must be negative because of the unit sum constraint. However, he was unable to propose a means to model such data in a way that removed the effect of the constraint.

1.3 Log and log-ratio transforms

The first step towards modern compositional data analysis was arguably the use by McAlister (1879) of Log-Normal distributions to model data that are constrained to lie in positive real space. Interestingly, he proposed this as the law of the geometric mean (versus the Normal distribution as the law of the arithmetic mean) and pointed out the lack of practical value for variance of a variable that must be positive, which can be seen in retrospect as recognition of the need for a different metric for data from restricted sample spaces, that takes constraints into account. Instead, he emphasized the meaning of the cumulative distribution. This is by no means the only way to model data on the positive real line and competes with, for example, the Gamma and Weibull distributions. It is equivalent to taking a log transform of the data, so that the non-negative constraint is removed, and then assuming a Normal distribution. One of the key texts for the Log-Normal distribution is the book by Aitchison and Brown (1969). However, this only addresses the non-negative constraint of compositional data and does not address the unit sum constraint.

The simplest meaningful example of a composition is with just two components, so the unit-sum constraint implies that the second component is just one minus the first component. This is just the situation that arises with probabilities for a binary outcome. Cox and Snell (1989) use the logit or logistic transformation of the probability in this case, which enables the use of regression models for the logit transformed probabilities. However, it appears that nobody saw the potential for a similar approach for the more general case of compositional data until the first known reference to using the log-ratio transform to solve the constraint problem for compositional (or simplicial) data by Obenchain in a personal communication to Johnson and Kotz (Kotz *et al.* 2000). Indeed, Obenchain contributed to the discussion of the Royal Statistical Society paper by Aitchison (1982), where he stated that he became discouraged by the problem of zero components and thus never attempted to publish his simplex work, even though he had derived many properties of the logistic-normal distribution.

The first public introduction of the properties of the logistic-normal distribution can be found in Aitchison and Shen (1980). This distribution is written in terms of log-ratios relative to the last component, so that $\mathbf{y}(\mathbf{x}) = \{\log(x_1/x_D), \ldots, \log(x_{D-1}/x_D)\}$ follows a Multivariate Normal distribution.

Up to that time, the only known tractable distribution on the simplex was the Dirichlet distribution. However, the Dirichlet distribution has some very restrictive properties, such

as complete subcompositional independence, i.e. for each possible partition of the composition, the set of all its subcompositions must be independent. This makes it impossible to model any reasonable dependence structure for compositional data using the Dirichlet distribution. In contrast, the logistic-normal distribution yields a distribution on the interior of the simplex that does not require these inflexible properties, but instead they become testable linear hypotheses on the covariance matrix within a broad flexible modelling framework. In addition, the Aitchison and Shen (1980) paper showed that the logistic-normal distribution is close to any Dirichlet distribution in terms of the Kullback–Leibler divergence. Later Aitchison (1985) derived a more general distribution that contains both the Dirichlet and logistic-normal distributions, although the potential for using this distribution for testing Dirichlet against logistic-normal distributions within the same class is diminished as these hypotheses are on the boundary of the parameter space. More recently, the generalization of the logistic-normal distribution to the additive logistic skew-normal distribution on the simplex (Mateu-Figueras *et al.* 2005) applies the skew-normal distribution (Azzalini 2005) to log-ratios on the simplex and offers the useful possibility of modelling data where the distribution of $\mathbf{y}(\mathbf{x})$ is not symmetrical. Use of the logistic-normal distribution opens up the full range of linear modelling available for the multivariate Normal distribution in $\mathbb{R}^D$.

1.4 Subcompositional dependence

As mentioned above, the logistic-normal distribution has the ability to model useful dependence structures. In his seminal book, Aitchison (1986) developed this idea, showing that the covariance structure can be modelled in terms of covariances on the log scale and is completely determined by the $D(D-1)/2$ log-ratio variances $\tau_{ij} = \mathrm{Var}\{\log(x_i/x_j)\}$(where $i = 1, \ldots, D-1; j = i+1, \ldots, D$).

However, finding a convenient matrix formulation seems tricky, either yielding formulations that require selecting a specific component as divisor [when using $\mathbf{\Sigma}$, which is the log-ratio covariance matrix for the $D-1$ log-ratios relative to one component as divisor (Aitchison 1986, p. 77)], have a zero diagonal [when using $\mathbf{T}$, which is the variation matrix for all pairs of log-ratios (Aitchison 1986, p. 76)] or are singular [when using $\mathbf{\Gamma}$, which is the centred log-ratio covariance matrix (Aitchison 1986, p. 79)]. However, it turns out that there are simple linear relationships between these alternative formulations, so it is feasible to choose whichever formulation is simplest to use in any specific context.

1.5 alr, clr, ilr: which transformation to choose?

One key question for using the log-ratio transformations is choosing the divisor. Most of the literature initially used an arbitrary component as the divisor, known as using alr (additive log-ratio) transformation. This is potentially problematic because the distances between points in the transformed space are not the same for different divisors. However, as shown in Aitchison (1986) and further developed in Aitchison *et al.* (2000), linear statistical methods with compositional data as the dependent variable are invariant to the choice of divisor as the implicit linear transformations between different representations cancel out in any F ratio of quadratic or bilinear forms, so this is a conceptual rather than practical problem.

One way of avoiding this problem of choosing a divisor is to divide by the geometric mean, known as the clr (centred log-ratio) transformation. As noted above, the disadvantage of this is that the clr covariance matrix is singular, making it difficult to use in some standard statistical procedures without adaption.

A key step forward was recognition that compositions can be represented by their coordinates in the simplex with a suitable orthonormal basis. This suggests an alternative transformation, known as ilr (isometric log-ratio) transformations (Egozcue *et al.* 2003), which avoids the arbitrariness of alr and the singularity of clr. Thus ilr has significant conceptual advantages, but unfortunately, there is no clear 'simplest' or canonical basis, unlike $\mathbb{R}^D$. One possibility is to use a sequential binary partition of the components (Egozcue and Pawlowsky-Glahn 2005), known as balances, although this alone still does not ensure uniqueness. This approach is explained in detail in Chapter 3. However, despite the mathematical elegance of this approach, it has practical disadvantages in the relative difficulty of choosing the basis when that is not motivated by the statistical question being investigated and also when relating the coordinates back to the original statistical question.

1.6 Principles, perturbations and back to the simplex

At this point, the reader may have concluded that compositional data analysis is entirely a pragmatic approach to avoiding the unit sum constraint that may have mathematical weaknesses. Indeed, mathematical geologists, typified by Rehder and Zier (2001) argued that log-ratio analysis implied an illogical and arbitrary distance metric. In fact, the log-ratio approach can be derived entirely from a few key principles, which enable the derivation of the entire mathematical framework including an appropriate distance metric on the simplex. As explained in Aitchison *et al.* (2000), it should be obvious that compositional data analysis can only make meaningful statements about ratios of components, i.e. the first principle is scale invariance. This should be obvious in that compositional data is unit-free, but some geologists, such as Watson and Philip (1989), did not find this obvious. The second key principle is subcompositional coherence (Aitchison 1992), which states that inferences about subcompositions should be consistent, regardless of whether the inference is based on the subcomposition or the full composition. For $\mathbb{R}^D$, this would translate into the self-evident principle that inference about a subset of variables should be the same regardless of whether we base the inference on the subset of variables or the full set. Any meaningful metric for the simplex should satisfy these two principles and the Euclidean metric for $\mathbb{R}^D$ clearly does not satisfy either for compositional data. Aitchison (1986) introduced the idea of perturbation as the analogue to linear operations in $\mathbb{R}^D$, which was further developed in Aitchison and Ng (2005). A perturbation $\mathbf{p} = (p_1, \ldots, p_D)$ is a differential scaling operator that when applied to the composition $\mathbf{x} = (x_1, \ldots, x_D)$ yields the composition

$$\mathbf{X} = \mathbf{p} \oplus \mathbf{x} = \mathcal{C}(p_1 x_1, \ldots, p_D x_D),$$

where $\mathcal{C}$ is the closure operator that scales elements to ensure that we remain in the unit simplex.

The set of perturbations (if restricted to $\mathcal{S}^D$) form a group with an inverse and an identity perturbation $\mathbf{e} = (1/D, \ldots, 1/D)$. As any composition can be expressed as a result of a perturbation on any other composition, the distance between any two compositions must be

expressible in terms of perturbations. Perturbation clearly corresponds to addition in $\mathbb{R}^D$ and we can define powering to correspond to multiplication in $\mathbb{R}^D$ as

$$\mathbf{X} = \mathbf{a} \odot \mathbf{x} = \mathcal{C}(x_1^a, \ldots, x_D^a).$$

The simplicial metric, or Aitchison distance is then given by

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{i=1}^{D} \left[\log \frac{x_i}{\mathrm{g_m}(\mathbf{x})} - \log \frac{y_i}{\mathrm{g_m}(\mathbf{y})} \right]^2 \right\}^{\frac{1}{2}},$$

where $\mathrm{g_m}(\cdot)$ is the geometric mean of the components, which can be shown to satisfy all the usual metric axioms and to depend only on perturbation distance. It is also easy to show that this metric satisfies the two key principles mentioned above.

The centre for a compositional distribution is then

$$\mathrm{Cen}[\mathbf{x}] = \mathcal{C}(\exp\{\mathrm{E}[\log(\mathbf{x})]\}),$$

with the variation matrix, **T**, as the most convenient measure of variability. In summary, this allows us to transfer the analysis back to the simplex, without the asymmetry of using alr.

1.7 Biplots and singular value decompositions

It is essential to have simple ways to summarize and display multivariate data sets. Fortunately, singular value decompositions and the related graphical tool of the biplot (Gabriel 1971), provide precisely the tools we need for compositional data when adapted to the simplex (Aitchison and Greenacre 2002). The biplot for the simplex is based on a singular value decomposition of the row and column centred log-ratio matrix. It enables us to graphically display which combinations of the log-ratios contain large and small amounts of variability. The former provides a useful simplification of the major contributions to total compositional variability, while the latter identifies any likely linear dependencies amongst the log-ratios.

1.8 Mixtures

One important application of compositional data is as the covariate that determines a mixture. This yields log-contrast models for experiments with fixed mixtures where the dependence is only on the composition (Aitchison and Bacon-Shone 1984).

Compositional data can also occur doubly as the mixture of compositions. In this case, the mixed composition does not stay within the class of logistic-normal distributions, but can often be approximated well by a logistic-normal distribution as shown in Aitchison and Bacon-Shone (1999).

One specific problem where the mixture of compositions occurs is what geologists call the end-member problem (Renner 1993; Weltje 1997). In this case, the key question is which of the end members (of usually known compositions) are being mixed to form the

outcome composition. A full Bayesian analysis of the end-member problem including spatial dependence is found in Palmer and Douglas (2008).

1.9 Discrete compositions

In the discussion of the Royal Statistical Society paper by Aitchison (1982), R.L. Plackett raised the question about how best to model discrete compositions and whether this might provide a solution to the problem of zeros (see below). The first full analysis of discrete data using compositional data models can be found in Billheimer *et al.* (2001), who use the logistic-normal distribution to model the probabilities for a multinomial distribution to allow much more sophisticated modelling of the occurrence data for species. This can be seen as a more sophisticated approach to the multivariate count modelling of Aitchison and Ho (1989), who use a logistic-normal model for the log means of Poisson data. However, the resultant data in both cases were counts, rather than discrete compositions, although in Billheimer's case, it is the relative occurrence that is of interest. As shown in Bacon-Shone (2008), it may be helpful to model discrete compositions even without knowing the total counts, helped by the knowledge that the original counts are non-negative integers.

1.10 Compositional processes

One of the nice consequences of analysing in the simplex, is that it is easy to investigate compositional processes as in Thomas and Aitchison (2005), who examine dependence on time through

$$D\mathbf{x}(t) = \mathcal{C}\left\{\exp\left[\frac{\mathrm{d}}{\mathrm{d}t}\log\mathbf{x}(t)\right]\right\},$$

where $\mathrm{d}/\mathrm{d}t$ denotes differentiation with respect to time.

This approach allows easy investigation and modelling of any processes that can be parameterized, including the possibility of change-points (Bacon-Shone 2011).

1.11 Structural, counting and rounded zeros

Aitchison recognized from the start that there was a need to solve the problem of zeros in compositional data as the log-ratio is undefined in this case. He wrote a much earlier paper (Aitchison 1955) which looked at the related problem of zeros for non-negative data, which presents a similar problem when using Log-Normal, Gamma or Weibull distributions, all of which have zero probability of observing zero. This paper used a conditional approach that separates the zeros from the continuous distribution and was applied to household expenditure data, which is a compositional data problem when classified into different categories of expenditure.

The original approach to compositional zeros in Aitchison (1982) was to simply replace all zeros by a small positive amount less than the detection limit, with the closure operator applied to apply the unit sum constraint and then sensitivity analysis to check the impact of

the replacement value. However, this positive replacement approach potentially distorts the compositional data.

Proper 'in the simplex' approaches were independently proposed by Martín-Fernández *et al.* (2000) and Fry *et al.* (2000). Palarea-Albaladejo *et al.* (2007) used a parametric model to handle the zeros as missing data using the expectation-maximization (EM) algorithm.

As pointed out by Aitchison and Kay (2003) and by Bacon-Shone (2003), zeros can occur for at least three distinct reasons. First, there may be a structural reason why the component must be zero, such as alcohol expenditure components for household expenditure data in families that do not drink alcohol. This situation is best modelled by the conditional approach. Secondly, there may be a zero because of an underlying discrete process (Bacon-Shone 2008), such as expenditure on white goods (i.e. major household appliances) in household expenditure data, where people may go several years between making purchases and may miss capture in the data collection process. This situation is best modelled by modelling the underlying discrete process. Lastly, there may be a limit in the measurement or recording processes, such that very small components are recorded as zero. For this situation, the approaches of Martín-Fernández *et al.* (2000) and Fry *et al.* (2000) mentioned above seem most relevant.

Recently Butler and Glasbey (2008) proposed another modelling approach to compositional data with zeros, using Euclidean projections onto the simplex, with the probability 'outside' the simplex used to model the point probabilities on the boundaries. Unfortunately, this approach does not have a special case of log-ratio analysis and fails the test of the two key principles mentioned above.

A more comprehensive discussion of how to handle zeros in compositions can be found in Chapter 4.

1.12 Conclusion

This brief summary shows how much progress has been made in the last century in finding appropriate analyses for compositional data, much of it in the last 30 years, relying heavily on the insights of John Aitchison.

Acknowledgement

This research has been partially supported by the Research Grants Council, Hong Kong (Grant HKU 700303).

References

Aitchison J 1955 On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association* **50**(271), 901–908.

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 1985 A general class of distributions on the simplex. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **47**(1), 136–146.

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

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

Aitchison J and Bacon-Shone J 1984 Log contrast models for experiments with mixtures. *Biometrika* **71**(2), 323–330.

Aitchison J and Bacon-Shone J 1999 Log contrast models for experiments with mixtures. *Biometrika* **86**(2), 351–364.

Aitchison J and Brown JAC 1969 *The Lognormal Distribution with Special Reference to its Uses in Econometrics*. Department of Applied Economics Monograph: 5. Cambridge University Press, Cambridge (UK). 176 p.

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

Aitchison J and Greenacre M 2002 Biplots for compositional data. *Applied Statistics* **51**(4), 375–392.

Aitchison J and Ho C 1989 The multivariate Poisson-log normal distribution. *Biometrika* **76**(4), 643–653.

Aitchison J and Kay J 2003 Possible solution of some essential zero problems in compositional data analysis. In *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop* (ed. Thió-Henestrosa S and Martín-Fernández JA). http://ima.udg.es/Activitats/CoDaWork03/. University of Girona, Girona (Spain). CD-ROM.

Aitchison J and Ng K 2005 The role of perturbation in compositional data analysis. *Statistical Modelling* **5**(2), 173–185.

Aitchison J and Shen SM 1980 Logistic-normal distributions. Some properties and uses. *Biometrika* **67**(2), 261–272.

Aitchison J, Barceló-Vidal C, Martín-Fernández JA and Pawlowsky-Glahn V 2000 Logratio analysis and compositional distance. *Mathematical Geology* **32**(3), 271–275.

Azzalini A 2005 The skew normal distribution and related multivariate families. *Scandinavian Journal of Statistics* **32**(2), 159–188.

Bacon-Shone J 2003 Modelling structural zeros in compositional data. In *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop* (ed. Thió-Henestrosa S and Martín-Fernández JA). http://ima.udg.es/Activitats/CoDaWork03/. University of Girona, Girona (Spain). CD-ROM.

Bacon-Shone J 2008 Discrete and continuous compositions. In *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop* (ed. Daunis-i Estadella J and Martín-Fernández J), p. http://hdl.handle.net/10256/723. University of Girona, Girona (Spain). 11 p.

Bacon-Shone J 2011 Mixing of compositions at points and along lines. *Computers & Geosciences* **37**(5), 692–695.

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

Butler A and Glasbey C 2008 A latent gaussian model for compositional data with zeros. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **57**(5), 505–520.

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

Cox D and Snell E 1989 *Analysis of Binary Data*, 2nd edition. Chapman and Hall/CRC, London (UK). p. 236.

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

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

Fry JM, Fry TRL and McLaren KR 2000 Compositional data analysis and zeros in micro data.. *Applied Economics* **32**(8), 953–959.

Gabriel KR 1971 The biplot – graphic display of matrices with application to principal component analysis. *Biometrika* **58**(3), 453–467.

Kotz S, Balakrishnan N and Johnson NL 2000 *Continuous Multivariate Distributions. Volume I, Models and Applications*. Wiley Series in Probability and Statistics. Wiley-Interscience, New York, NY (USA). 730 p.

Martín-Fernández JA, Barceló-Vidal C and Pawlowsky-Glahn V 2000 Zero replacement in compositional data sets. In *Studies in Classification, Data Analysis, and Knowledge Organization. Proceedings of the 7th Conference of the International Federation of Classification Societies (IFCS'2000)* (ed. Kiers H, Rasson J, Groenen P and Shader M). Springer-Verlag, Berlin (Germany) pp. 155–160.

Mateu-Figueras G, Pawlowsky-Glahn V and Barceló-Vidal C 2005 The additive logistic skew-normal distribution on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* **19**(3), 205–214.

McAlister D 1879 The law of the geometric mean. *Proceedings of the Royal Society of London* **29**, 367–376.

Palarea-Albaladejo J, Martín-Fernández JA and Gómez-García JA 2007 Parametric approach for dealing with compositional rounded zeros. *Mathematical Geology* **39**(7), 625–645.

Palmer MJ and Douglas GB 2008 A bayesian statistical model for end member analysis of sediment geochemistry, incorporating spatial dependences. *Journal of the Royal Statistical Society. Series C: Applied Statistics* **57**(3), 313–327.

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.

Rehder S and Zier U 2001 Letter to the Editor: Comment on 'Logratio analysis and compositional distance' by J. Aitchison, C. Barceló-Vidal, J.A. Martín-Fernández and V. Pawlowsky-Glahn. *Mathematical Geology* **33**(7), 845–848.

Renner RM 1993 The resolution of a compositional data set into mixtures of fixed source components. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **42**(4), 615–631.

Tanner J 1949 Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *Journal of Applied Physiology* **2**(1), 1–15.

Thomas CW and Aitchison J 2005 Compositional data analysis of geological variability and process: a case study. *Mathematical Geology* **37**(7), 753–772.

Watson DF and Philip GM 1989 Measures of variability for geological data. *Mathematical Geology* **21**(2), 233–254.

Weltje JG 1997 End-member modeling of compositional data: numerical-statistical algorithms for solving the explicit mixing problem. *Mathematical Geology* **29**(4), 503–549.

2

Basic concepts and procedures

Juan José Egozcue[1] and Vera Pawlowsky-Glahn[2]
[1] *Department of Applied Mathematics III, Technical University of Catalonia, Spain*
[2] *Department of Computer Science and Applied Mathematics, University of Girona, Spain*

2.1 Introduction

In all experimental fields, large amounts of compositional data (CoDa) can be found. They describe quantitatively the parts of some whole. They appear as proportions, percentages, concentrations, absolute and relative frequencies, spreading and distribution functions. Their units are also diverse; they range from percentages, parts per unit, or parts per million, to other non-closed units like molar concentrations or absolute frequencies. Often, the total amount is irrelevant or the analyst is not interested in it. For instance, in order to study the political and sociological framework of an election, the total number of electors per circumscription, representing the size of such a region, is considered external to the study, and only the proportions between the number of votes to candidates are considered. In this case, the proportions of votes to lists can be considered compositional, even if they are presented as absolute number of votes. In geology or biology, the mass of a material sample can be considered irrelevant if analysts are only interested in the geo-/biochemical composition of that sample. The analysis of data expressed as proportions carries a number of problems that have been studied for a long time in fields like geology and biology. One of the first examples comes from the field of biologic morphology and is authored by one of the founders of modern statistics: K. Pearson (1897). In geology, the study of CoDa was particularly intensive in the 1950s and 1960s (Chayes 1960). In biology, some attempts can be found (Mosimann 1962; Connor and Mosimann 1969). But the first consistent methodological proposal to deal with CoDa did not

Compositional Data Analysis: Theory and Applications, First Edition. Edited by Vera Pawlowsky-Glahn and Antonella Buccianti.

arrive until the 1980s. It was introduced by J. Aitchison (1982, 1986). The main point is the statistical analysis of log-ratios and the statement of the principles of CoDa analysis.

Despite the advantages offered by techniques based on log-ratios, they did not have the success one could expect, and many scientists continued (and continue) applying the traditional statistical methods without taking into account the compositional character of their data. At the beginning of the 2000s, several formal contributions were published (Billheimer *et al*. 2001; Pawlowsky-Glahn and Egozcue 2001; Aitchison *et al*. 2002), which allow a better systematic approach to the methods already proposed in the 1980s (Aitchison and Egozcue 2005) (see Chapter 1). Nowadays, CoDa analysis can be reduced to three steps: the representation of data in log-ratio type coordinates; the (traditional) analysis of the coordinates as real random variables; and the interpretation of resulting models either in coordinates, or expressing the results in terms of the original units. This method, termed *principle of working in coordinates* (see Chapter 3), is based on the invariance of the analysis under change of basis. However, these techniques are not yet widely used, but a growing interest is emerging in several scientific fields.

2.2 Election data and raw analysis

To illustrate the anomalous behaviour of standard statistical methods when applied to raw CoDa (Aitchison 1997) we use the provisional results of the November 2010 elections to the *Parlament de Catalunya* (parliament of Catalonia), an autonomous community of Spain. The votes have been recorded for 41 regions, which are a subdivision of the electoral provinces. The data set (Cat10) contains the number of electors (elect), the number of votes, including *none of the above* (nota), null votes (null) and valid votes to parties and coalitions. The difference of electors minus votes gives the abstention (abst). Table 2.1 shows the first,

Table 2.1 Three records (first, intermediate and last) of Cat10. Number of votes per region and categories: number of electors (elect); abstention (abst); none of the above (nota); null votes (null); parties and coalitions in the outgoing Catalan Parliament; votes to amalgamated lists of candidates (other).

	Alt Camp	Barcelonès	Vallès Oriental
elect	32027	1572425	283230
abst	13418	629865	111884
nota	483	28306	5106
null	173	5722	1009
C's	285	39194	6305
CiU	8183	317695	67887
ERC	1679	56200	11199
ICV	931	85579	13119
PSC	2856	188802	29738
PP	1556	139890	18757
other	2463	81172	18226

an intermediate and the last record of the data set by columns. Several minor parties and coalitions, one of them achieving four representatives in the parliament in these elections but not present in the outgoing parliament, have been amalgamated into a single category, called *other*; major parties or coalitions are *Convergència i Unió* (CiU), virtual winner of the 2010 elections, and *Partit dels Socialistes de Catalunya* (PSC); other parties represented in the outgoing parliament of Catalonia, are (in alphabetical order): *Ciutadans-Partido de la Ciudadanía* (C's), *Esquerra Republicana de Catalunya* (ERC), *Iniciativa per Catalunya Verds-Esquerra Unida i Alternativa* (ICV), and *Partit Popular* (PP).

Attention is focused on the inter-relations between all categories of votes (including abstention) and a possible dependence of the distribution of votes among the different parties on the total population of the region.

A first (fruitless) attempt to analyse the election data consists in a correlation study of the raw number of votes. Correlations obtained in this way are clearly dominated by the size of the region, represented by the number of electors, and all of them attain values larger than 0.9. Some of these correlations are shown in Table 2.2, case **A**. To filter out the influence of size dividing each number of votes by the total number of electors in a region is common practice. Effectively this removes the *size of region* effect. Denote this approach **B**, corresponding to per unit or, equivalently, percentage of vote for each category. The correlation matrix obtained for approach **B** differs completely from the correlation matrix of case **A**. The high positive correlations completely disappear. Also, some significant correlations are observed. But the fact that each record of data adds up to 100 when data are expressed in percentage, induces some negative correlations. This effect, called *negative bias*, appears systematically when analysing the correlation of *closed data*, i.e. of a data set with records (observations) which are vectors with positive components adding up to a constant (in this case 100). Some values are shown in Table 2.2, replacing the large positive correlations in case **A**, supporting the impression that the high correlations were a size effect.

It should be possible to reproduce the conclusions obtained by an analyst from correlations for case **B** – or they should be at least compatible with those obtained by another analyst – examining correlations when proportions of votes are expressed in a different way. For instance, approach **B** considers proportions of votes over the whole census of electors; approach **C** may consider only proportion of votes to parties and coalitions, thus excluding abstention, nota and null votes; approach **D** may define proportions of parties and coalitions

Table 2.2 Spurious correlation: Pearson's correlation coefficients between votes to parties or coalitions and the total number of electors depending on approaches **A**, **B**, **C**, **D**, **F**. See text for an explanation.

		elect C's	elect PSC	C's ERC	C's CiU	PSC C's	PSC PP	PSC ERC	PSC ICV
A	absolute vote	0.995	0.997	0.981	0.985	0.999	0.996	0.986	0.998
B	electors	0.668	0.282	−0.794	−0.792	0.392	0.471	−0.334	0.324
C	candidates	0.646	0.304	−0.797	−0.804	0.521	0.590	−0.456	0.407
D	in parliament	0.642	0.306	−0.805	−0.764	0.509	0.537	−0.525	0.372
E	minor	0.544		−0.784					
F	C's, PSC, PP	0.588	−0.362			−0.712	−0.955		

that were present in the outgoing parliament; option **E** studies the proportion of votes distributed among minor parties (C's, ERC, ICV, other); option **F** may be to study relationships between votes to two large parties (PSC, PP), well represented in the Spanish parliament, with a recently born C's that is supposed to take some votes from the former parties. Correlation matrices for all these situations have been computed. Approach **B** assigns each potential vote to a category and takes the proportions of such votes. **B** considers a complete composition. Situations **C**–**F**, correspond to analysing some *subcomposition* of the original one. As this is common practice when studying percentage data (closed data), one expects to find coherent results when analysing different subcompositions with common categories. Some results have been given in Table 2.2 for an easy comparison; they show some incoherent results. Even the correlations with an external variable (elect) change incoherently from approach to approach. This phenomenon is called *spurious correlation* (Pearson 1897; Aitchison 1986). It invalidates multivariate techniques based on covariances of the raw composition.

2.3 The compositional alternative

Unsuccessful experiences analysing CoDa, similar to the example presented in Section 2.2, have been accumulated over a century. Based on them, J. Aitchison (1982, 1986) stated some principles for the analysis of CoDa. They have been reformulated several times (Barceló-Vidal *et al.* 2001; Martín-Fernández *et al.* 2003; Aitchison and Egozcue 2005; Egozcue 2009) according to new theoretical developments. The first step is the definition of CoDa, which essentially coincides with that stated in Section 2.1 above: *compositional data quantitatively describe the parts of some whole and they provide only relative information between their components*. Note that CoDa appear as vectors of two or more positive components, although frequently one component is omitted as it is the difference of the shown components to the total. However, only ratios of components carry information. This leads to the following principles.

2.3.1 Scale invariance: vectors with proportional positive components represent the same composition

In other words, if a composition is scaled by a constant, e.g. changing from parts per unit to percentages, the information carried is completely equivalent. Accordingly, vectors of proportional positive components form an equivalence class. Therefore, it is natural to select a representative of the equivalence class to facilitate both the analysis and the interpretation. The traditional way to do that is to normalize the vector in such a way that the components sum to a given constant κ, which can be 1, 100, 1000, 10^6, or any other positive constant. This selection is formalized by the *closure* operation. For $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ a vector with D positive components, its *closure* is defined as

$$\mathcal{C}\mathbf{x} = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right). \tag{2.1}$$

The components of the closed vector are called *parts*, relative to a *total* κ. The set of vectors with D positive components summing to the constant κ form the D-part simplex, denoted by $\mathcal{S}^D$. The compositions equivalent to $\mathbf{x}$ are represented by $\mathcal{C}\mathbf{x}$.

For instance, the votes in Cat10 add up to the number of electors in each region. If we are not interested in the number of electors, which is a measure of the size of the region, the ratio of the number of votes obtained by each party or coalition over the total of electors gives a per-unit distribution or proportion of votes, and the size has been filtered out. Multiplying these values by 100 we get percentages, and the vector of per-units and the vector of percentages convey exactly the same information.

2.3.2 Subcompositional coherence: analyses concerning a subset of parts must not depend on other non-involved parts

A subcomposition is a subset of components or parts of a composition. The study of a subcomposition, requires that the results are not contradictory with those obtained from the full composition. The principle of coherence can be summarized as two criteria: (a) the principle of scale invariance should hold for any of the possible subcompositions thus implying preservation of ratios of parts; (b) if a distance or divergence is used to compare compositions, this distance or divergence should be grater than or equal to that obtained comparing the corresponding subcompositions (*subcompositional dominance*).

Coherence is more subtle than the principle of scale invariance. In the example of Cat10, it means that the analysis of the complete distribution of votes (situation **B**, including abstention, nota and null votes) should be coherent with situation **C**, where only votes to parties and coalitions are taken into account, or even coherent with situation **F**, where the subcomposition (C's, PSC, PP) is analysed. The reason is that the selection of a subcomposition does not change the ratios between the parts and, since these ratios are the only information considered, the analysis should remain invariant when using the same parts from the composition and the subcomposition. Correlations shown in Table 2.2 depend on the subcomposition considered; they are clear examples of violation of the principle of subcompositional coherence: a measure of association between two parts should not depend on which other parts are in the subcomposition. See extreme cases in Table 2.2 where correlation between PSC and PP takes values as diverse as 0.590 (case **C**) and −0.955 (case **F**).

Subcompositional dominance calls for a way to measure distances between compositions and subcompositions which follows the rule of a projection: distances become smaller in a projection. It is logical to ask if the ordinary Euclidean distance between real vectors can be used. This is not the case, as both the principle of scale invariance and of subcompositional dominance are violated. In fact, if two vectors with positive components are multiplied by a positive constant c, then the Euclidean distance between them is multiplied by c, violating the principle of scale invariance. Subcompositional dominance is also violated by the ordinary Euclidean distance between compositional vectors. This fact is illustrated in Table 2.3. Euclidean distances between compositions of votes in two couples of regions are

Table 2.3 Euclidean distances between regions 1 and 2 measured on subcompositions corresponding to situations **B**, **C**, **D**, **F**.

Region 1	Region 2	**B**	**C**	**D**	**F**
Barcelonès	Alt Camp	0.087	0.155	0.178	0.116
Barcelonès	Vallès Oriental	0.049	0.083	0.093	0.049

considered: Barcelonès-Alt Camp and Barcelonès-Vallès Oriental. The raw number of votes for these regions are shown in Table 2.1. Consider the sequence of subcompositions corresponding to situations **B**, **C**, **D**, **F** defined in Section 2.2. As the corresponding subcompositions are nested from right to left in Table 2.3, subcompositional dominance would expect decreasing distances from left to right. This is not the case and, therefore, Euclidean distance for compositions is inappropriate for CoDa analysis (Palarea-Albaladejo *et al*. 2011).

2.3.3 Permutation invariance: the conclusions of a compositional analysis should not depend on the order of the parts

This is obvious in many cases. For example, in geochemical compositions it is quite frequent to record the parts in alphabetical order. The same happens in the example Cat10. But in some cases the parts can be assumed to be ordered. A typical example is the grain size distribution of a sediment: the particles are classified, after sieving, into size categories. Applying a compositional analysis, the information due to the order of the different classes, plays no role.

2.4 Geometric settings

To satisfy the requirements described in Section 2.3, a geometry of the D-part simplex, $\mathcal{S}^D$, is required. The development of the concepts suggested by Aitchison (1986) has lead to the *Aitchison geometry* of the simplex (Pawlowsky-Glahn and Egozcue 2001). This geometry, being Euclidean, requires specific definitions and a specific metric.

Consider the compositions $\mathbf{x}$, $\mathbf{y} \in \mathcal{S}^D$. The *perturbation* of $\mathbf{x}$ with $\mathbf{y}$ is defined as the composition

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D), \tag{2.2}$$

and *powering* of $\mathbf{x}$ by a real number α as the composition

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha). \tag{2.3}$$

It is easy to show that for $\mathbf{n} = \mathcal{C}(1, 1, \ldots, 1)$ it holds $\mathbf{x} \oplus \mathbf{n} = \mathbf{x}$. Thus, the composition with all equal parts is the neutral element of perturbation. *Perturbation* and *powering*, defined in $\mathcal{S}^D$, satisfy the requirements for operations of a vector space. But the main advantage of perturbation is that, in addition to satisfying the principles of compositional analysis, it has usually an interpretation in the field analysed.

The usefulness of perturbation and powering is illustrated using the example Cat10 and the regions Barcelonès, Alt Camp and Vallès Oriental. If interest lies, e.g., on the change of the distribution of votes from one region to another, a simple model is that the composition of votes in any region (Alt Camp or Vallès Oriental), $\mathbf{y}_i$, can be obtained from the composition in Barcelonès, $\mathbf{x}$, perturbed by some composition $\mathbf{p}_i$ that describes the change as a compositional shift in the simplex. Note that perturbation $\mathbf{p}_i$ can also be expressed as a difference perturbation denoted with the symbol $\ominus$:

$$\mathbf{y}_i = \mathbf{x} \oplus \mathbf{p}_i, \mathbf{p}_i = \mathbf{y}_i \ominus \mathbf{x} = \mathbf{y}_i \oplus ((-1) \odot \mathbf{x}). \tag{2.4}$$

Table 2.4 Modelling change between regions. See text for details.

	$\mathbf{x}$ %	$\mathbf{y}_1$ %	$\mathbf{y}_2$ %	$\mathbf{p}_1$ %	$\mathbf{p}_2$ %	$\mathbf{p}_1$ factor	$\mathbf{p}_2$ factor	$\mathbf{p}_1$ % change	$\mathbf{p}_2$ % change
abst	40.1	41.9	39.5	10.7	10.0	1.046	0.986	4.6	−1.4
nota	1.8	1.5	1.8	8.6	10.1	0.838	1.001	−16.2	0.1
null	0.4	0.5	0.4	15.2	9.9	1.484	0.979	48.4	−2.1
C's	2.5	0.9	2.2	3.7	9.0	0.357	0.893	−64.3	−10.7
CiU	20.2	25.6	24.0	12.9	12.0	1.265	1.186	26.5	18.6
ERC	3.6	5.2	4.0	15.0	11.2	1.467	1.106	46.7	10.6
ICV	5.4	2.9	4.6	5.5	8.6	0.534	0.851	−46.6	−14.9
PSC	12.0	8.9	10.5	7.6	8.9	0.743	0.874	−25.7	−12.6
PP	8.9	4.9	6.6	5.6	7.5	0.546	0.744	−45.4	−25.6
other	5.2	7.7	6.4	15.2	12.6	1.490	1.247	49.0	24.7
A. dist.				1.51	0.48				

Table 2.4 shows the perturbation from Barcelonès to Alt Camp, $\mathbf{p}_1$, and to Vallès Oriental, $\mathbf{p}_2$, in three different versions: as a composition expressed in percentage (%); as multiplicative (non-closed) factors; and as percentage of increase/decrease, which is a traditional form of presenting perturbations. For instance, a factor of 1.046 in abst from Barcelonès to Alt Camp means a 4.6% increment of abstention; a factor of 0.357 for C's, corresponds to a 64.3% decrease from Barcelonès to Alt Camp. The Alt Camp region, a rural area compared with Barcelona and its surroundings, shows substantial increments of votes to Catalan nationalist parties (CiU, ERC, other), a moderate increase of abstention, and a heavy null vote increase with respect to Barcelonès. Vallès Oriental is an industrial and populated area. Accordingly, it may have similarities with Barcelonès, which is mainly occupied by Barcelona, a crowded city, with a lot of services and partially industrial. In fact, although CiU, ERC and other also increase with respect to Barcelonès, the increase is not as large as observed for Alt Camp. A remarkable fact is that the abstention and null votes in Vallès Oriental are less than in Barcelonès, while nota votes are similar. The global size or magnitude of a perturbation is its Aitchison norm which is actually the Aitchison distance between the regions. The last row in Table 2.4 gives Aitchison distances (A. dist.) for both couples of regions, as discussed below.

The change of units in some or all the parts of a composition can also be viewed as a perturbation. Typical examples are found in chemistry, when concentrations in parts per million (ppm) of weight are changed to molar concentrations (Buccianti and Pawlowsky-Glahn 2005). This is done by multiplying each component by the inverse of the molar weight. Closing the resulting composition may be unnecessary in many cases. Still, it retains its compositional character. For the example Cat10, we can imagine some meaningful change of units. For instance, considering the subcomposition of candidate parties and coalitions, each one has invested some money in the campaign and, thus, a per-capita (per-elector) amount can be computed for each list. The votes can then be viewed as a return of the investment. The vector of these returns, expressed in monetary units, is still compositional, although closing it to percentages does not have a clear meaning.

Scale invariance required for a compositional analysis leads to the use of ratios of parts so that scale constants are cancelled. Furthermore, ratios can be considered in a relative scale

and taking their logarithms is then a natural choice. The analysis of CoDa is essentially based on the statistical analysis of log-ratios of parts. The simplest log-ratios are those comparing two parts. In Cat10, ln(CiU/PSC) or ln(abst/nota) are two of these simple log-ratios. More complex log-ratios can be useful in the analysis, but they must be scale invariant. Scale invariant log-ratios are called log-contrasts (Aitchison 1986) and are defined as

$$\ln\left(\prod_{i=1}^{D} x_i^{\alpha_i}\right) = \sum_{i=1}^{D} \alpha_i \ln(x_i), \quad \sum_{i=1}^{D} \alpha_i = 0, \tag{2.5}$$

where the condition on the coefficients α_i guarantees scale invariance.

Frequently, some questions can be answered analysing an appropriate log-contrast. The choice of the log-contrast depends on the stated problem and the interpretation of the composition. In the Cat10 context we can ask ourselves whether the log-ratio between parties declared as Catalan nationalists (CiU, ERC, other) and parties declared as opposite to Catalan nationalism (C's, PP) depends on the total number of electors of the region. A choice of the log-contrast is

$$z = \sqrt{\frac{6}{5}} \ln \frac{(\text{CiU} \cdot \text{ERC} \cdot \text{other})^{1/3}}{(\text{C's} \cdot \text{PP})^{1/2}},$$

where the coefficient in the square root has been added for normalization, as described below when defining balances. The log-contrast z can be computed for all regions and then it is correlated with the logarithm of the number of electors (elect). After removing one outlier (Vall d'Aran, a valley in the Pyrenees), the estimated correlation coefficient is -0.778 and the p-value for the F-test is less than 10^{-6}. Figure 2.1 shows ln(elect) and the log-contrast z

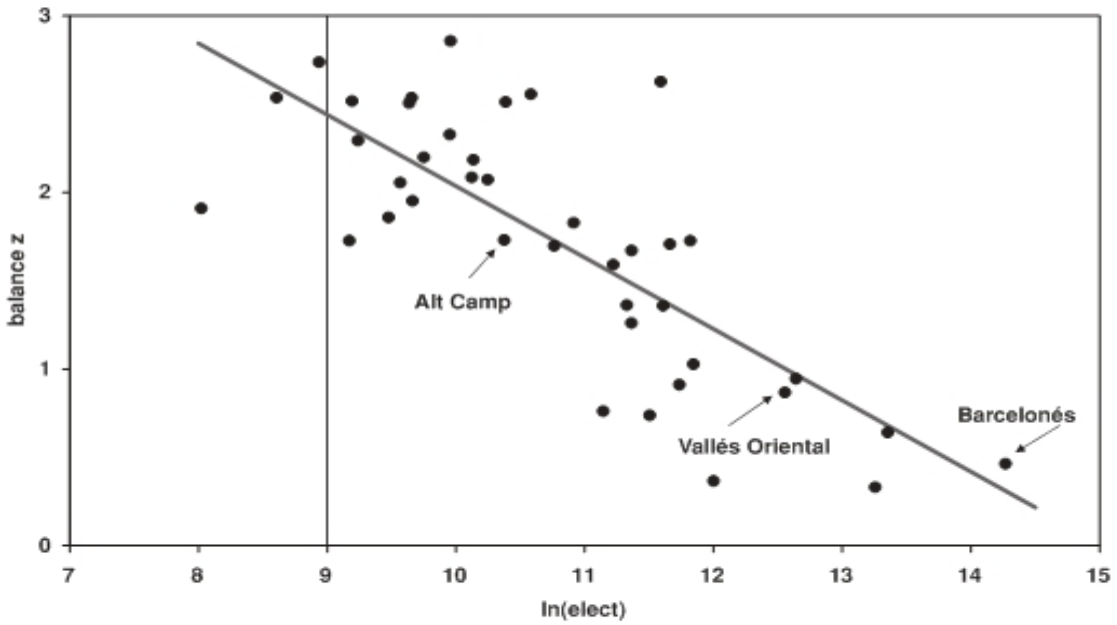

Figure 2.1 Regions considered in Cat10 characterized by the logarithm of the number of electors and the log-contrast z. The correlation coefficient is -0.778.

for each region and the linear model fitted. The significant negative correlation illustrated in Figure 2.1 allows an interpretation. Roughly speaking, the smaller logarithm of the number of electors, the more rural characteristics of the region, and vice versa the more industrial and economically active regions have a larger number of electors. The fitted model reflects the fact that the Catalan nationalist vote attains higher values in rural areas than in urban and industrial areas.

The previous example of regression of a log-contrast on an external variable illustrates the fact that certain aspects of compositions can be analysed using the appropriate log-contrast. Proper and complete representations of a composition using a set of log-contrasts were proposed in the 1980s (Aitchison 1986), so that all the information from the composition is reverted into the set of log-ratios. A first choice of these representations was the *additive-log-ratio* transformation (alr). If $\mathbf{x}$ is a composition in the D-part simplex $\mathcal{S}^D$,

$$\operatorname{alr}(\mathbf{x}) = \ln\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D}\right), \tag{2.6}$$

where the natural logarithm ln is applied componentwise. Consequently, the ith component is the simple log-ratio $\operatorname{alr}_i(\mathbf{x}) = \ln(x_i/x_D)$. The alr transformation is easily inverted to get the composition from the $D-1$ alr components and also reduces perturbation and powering to ordinary operations in the $D-1$ dimensional real space:

$$\operatorname{alr}((\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})) = \alpha \cdot \operatorname{alr}(\mathbf{x}) + \beta \cdot \operatorname{alr}(\mathbf{y}),$$

for any compositions $\mathbf{x}, \mathbf{y}$ and any real constants α and β. However, alr has the inconvenient of not being invariant under permutation of components and some statistical procedures may fail. In fact, we favour the Dth part (or any other part selected) which is present in all denominators of the log-ratios. To avoid these problems, Aitchison (1986) introduced the *centered log-ratio* transformation (clr), which represents a D-part composition using D clr coefficients. It is defined as

$$\mathbf{v} = \operatorname{clr}(\mathbf{x}) = \ln\left[\frac{x_1}{\mathrm{g_m}(\mathbf{x})}, \frac{x_2}{\mathrm{g_m}(\mathbf{x})}, \ldots, \frac{x_D}{\mathrm{g_m}(\mathbf{x})}\right], \quad \mathrm{g_m}(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}, \tag{2.7}$$

where the D coefficients $\operatorname{clr}_i(\mathbf{x}) = \ln(x_i/\mathrm{g_m}(\mathbf{x}))$ are log-contrasts [see Equation (2.5)]. From $\operatorname{clr}(\mathbf{x})$, the composition $\mathbf{x}$ is recovered with the inverse clr transformation

$$\mathbf{x} = \operatorname{clr}^{-1}(\mathbf{v}) = \mathcal{C}\exp(\mathbf{v}), \tag{2.8}$$

where the exponential function is applied componentwise to $\mathbf{v} = \operatorname{clr}(\mathbf{x})$. Similarly to the alr transformation, perturbation and powering in $\mathcal{S}^D$ correspond to sum and product in D-dimensional real space $\mathbb{R}^D$, i.e.

$$\operatorname{clr}((\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})) = \alpha \cdot \operatorname{clr}(\mathbf{x}) + \beta \cdot \operatorname{clr}(\mathbf{y}).$$

The drawback of the clr transformation is that it uses D coefficients, adding to zero, to represent a composition which has only $D-1$ free components, the dimension of $\mathcal{S}^D$. Moreover, the clr components change when working with a subcomposition.

The clr representation of compositions can be used to define a metric structure in the simplex. The Aitchison inner product, norm and distance for compositions in $\mathcal{S}^D$ are

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{y}) \rangle, \tag{2.9}$$

$$\|\mathbf{x}\|_a = \|\text{clr}(\mathbf{x})\|, \quad \text{d}_a(\mathbf{x}, \mathbf{y}) = \text{d}(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{y})), \tag{2.10}$$

where $\langle \cdot, \cdot \rangle$, $\| \cdot \|$, $\text{d}(\cdot, \cdot)$, denote the ordinary Euclidean inner product, norm and distance in $\mathbb{R}^D$. For instance, the Aitchison distance is

$$\text{d}_a(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{D} [\text{clr}_i(\mathbf{x}) - \text{clr}_i(\mathbf{y})]^2}.$$

Table 2.4, in its last row, shows the Aitchison distance between Barcelonès and the other two regions (Alt Camp, Vallès Oriental) in example Cat10. The first one is about three times the second, thus reflecting a moderate socio-economical proximity of Barcelonès and Vallès Oriental compared with the large difference between Barcelonès and Alt Camp. Aitchison distances between regions provide a tool for further multivariate analysis, e.g. cluster analysis or multidimensional scaling.

The Aitchison inner product, norm and distance honour the principles of compositional analysis (Section 2.3) and are therefore tools for a compositional analysis free of inconsistencies. Jointly with perturbation and powering, they provide an Euclidean structure to the simplex, called Aitchison simplicial geometry. This suggests to exploit the well-known properties of Euclidean spaces to analyse compositions: orthonormal basis, (orthonormal) coordinate representation, orthogonal projections, definitions of angles, ellipses, etc. An important step to use these concepts is to build orthonormal bases and their corresponding coordinates.

An orthonormal basis of $\mathcal{S}^D$ is a set of compositions $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}$ such that $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_a = 0$ for $i \neq j$, and $\|\mathbf{e}_i\|_a = 1$. For a fixed basis the coordinates of a composition are obtained using the function

$$\mathbf{x}^* = \text{ilr}(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a), \tag{2.11}$$

with inverse,

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \bigoplus_{j=1}^{D-1} x_j^* \odot \mathbf{e}_j. \tag{2.12}$$

The construction of orthonormal coordinates has been called *isometric log-ratio transformation* (ilr) (Egozcue *et al*. 2003) because the coordinates $x_j^* = \text{ilr}_j(\mathbf{x})$ are log-contrasts and are isometric:

$$\text{ilr}((\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{y})) = \alpha \cdot \text{ilr}(\mathbf{x}) + \beta \cdot \text{ilr}(\mathbf{y}), \tag{2.13}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{y}) \rangle, \tag{2.14}$$

$$\|\mathbf{x}\|_a = \|\text{ilr}(\mathbf{x})\|, \tag{2.15}$$

$$\text{d}_a(\mathbf{x}, \mathbf{y}) = \text{d}(\text{ilr}(\mathbf{x}), \text{ilr}(\mathbf{y})), \tag{2.16}$$

analogous to the properties given in (2.9) and (2.10) for the clr transformation. The difference is that the inner product, norm and distance between vectors of ilr coordinates correspond to $D-1$ dimensional real space, which is isomorphic to $\mathcal{S}^D$.

As in any other Euclidean space, an infinite number of orthonormal bases exist in $\mathcal{S}^D$. A simple way to represent compositions in ilr coordinates is to perform a *Singular Value Decomposition* (SVD) of the matrix of a clr transformed sample, an operation which is usually done to obtain a *biplot*, as described in Section 2.5. But the interpretation of the resulting log-contrasts [Equation (2.5)] can be difficult. Therefore, it can be convenient to build log-contrasts which have an interpretation adequate to the problem studied and, in particular, orthonormal coordinates defined by the analyst according to the problem he or she is trying to solve. One technique for doing so is based on a *Sequential Binary Partition* (SBP) of the parts of the composition (Egozcue and Pawlowsky-Glahn 2005a, 2006; Pawlowsky-Glahn and Egozcue 2006; Thió-Henestrosa *et al*. 2008). Each step of the partition, of a total of $D-1$ steps, gives rise to an ilr coordinate, now called *balance*, which is usually easy to interpret. Table 2.6 illustrates the SBP process. In a first step, SBP consists of dividing the composition into two groups of parts which are indicated by $+1$ and -1, as shown in the first row of Table 2.6. In further steps, each previously obtained group of parts is again subdivided into two groups until all groups are made of a single part. The first step in Table 2.6 consists of separating votes to parties and coalitions (-1) from abstention and nota or null votes $(+1)$. The second step separates abstention $(+1)$ from nota and null votes (-1). The fifth step separates declared Catalan nationalist coalitions $(+1)$ from parties which are present in the whole of Spain (-1), etc. Each step in the SBP is associated with one element of an orthonormal basis and one ilr coordinate. These ilr coordinates are called *balances* due to their peculiar form. For the jth row of the SBP matrix (Table 2.6) denote by $\mathbf{x}_+$ the group of r parts marked with a $+1$ and by $\mathbf{x}_-$ the group of s parts marked with a -1; then, the balance is

$$b_j = \sqrt{\frac{rs}{r+s}}\,\ln\frac{g_m(\mathbf{x}_+)}{g_m(\mathbf{x}_-)},$$

where $g_m(\cdot)$ is the geometric mean of its arguments. Balances are unit-norm orthogonal log-contrasts (Aitchison 1986) and they have a relatively easy interpretation (Egozcue and Pawlowsky-Glahn 2005b), as they are log-ratios of geometric means of groups of parts.

2.5 Centre and variability

Statistics synthesizes information from a sample using simple descriptors. The mean and variance-covariance are the most popular in multivariate scenarios. When dealing with CoDa, the geometry of their sample space $\mathcal{S}^D$ must be taken into account and, particularly, the Aitchison distance. Following Pawlowsky-Glahn and Egozcue (2001), consider a random composition $\mathbf{X}$ in $\mathcal{S}^D$ and define variability of $\mathbf{X}$ with respect to a composition $\mathbf{z}\in\mathcal{S}^D$ as $\mathrm{Var}(\mathbf{X},\mathbf{z}) = \mathrm{E}[\mathrm{d}_a^2(\mathbf{X},\mathbf{z})]$, where E and Var denote the ordinary expectation and variance in real space. The composition $\mathbf{z}$ minimizing $\mathrm{Var}(\mathbf{X},\mathbf{z})$ is called the centre of $\mathbf{X}$, and the minimum variability attained is the total variance. The centre of $\mathbf{X}$ is then expressed as

$$\mathrm{Cen}[\mathbf{X}] = \mathrm{ilr}^{-1}\{\mathrm{E}[\mathrm{ilr}(\mathbf{X})]\} = \mathcal{C}\exp(\mathrm{E}[\ln\mathbf{X}]). \tag{2.17}$$

The right-hand side of the equation can be used to define the centre as a *closed geometric mean* (Aitchison 1997). The centre Cen[**X**] plays the role of the multivariate mean when the sample space is the simplex.

The total variance can be expressed in three different ways, each of them providing a decomposition of total variance:

$$\text{totVar}[\mathbf{X}] = \frac{1}{D}\sum_{i=1}^{D-1}\sum_{j=i+1}^{D}\text{Var}\left[\ln\frac{X_i}{X_j}\right] \tag{2.18}$$

$$= \sum_{i=1}^{D}\text{Var}[\text{clr}_i(\mathbf{X})] \tag{2.19}$$

$$= \sum_{j=1}^{D-1}\text{Var}[\text{ilr}_j(\mathbf{X})]. \tag{2.20}$$

The first one (2.18) proposes the variance of all simple log-ratios as components, with the advantage of easy interpretation of each component. The decomposition into ilr components of variance (2.20) is globally more understandable because of the orthogonality of the ilr coordinates, while each individual coordinate may require an additional effort of interpretation. The centre, as well as the total variance and its components, can be estimated using the ilr components; their properties correspond to those of the estimators of the mean and the variance-covariance in real sample spaces (Pawlowsky-Glahn and Egozcue 2001, 2002).

A typical way to present estimations of the centre and the variability is the variation array (Aitchison 1986). Table 2.5 shows the variation array for Cat10 data. The lower triangle shows the sample mean value of the simple log-ratio of the corresponding two parts (numerator by column, denominator by row). Therefore, the last row of the array is the alr transformation, using *other* in the denominator, of the estimated centre. For instance, the mean log-ratio between CiU and PSC is positive (1.06) and indicates that, in mean, CiU gets more votes than PSC.

The upper triangle contains the sample variances of the same log-ratios. They are the sample version of the terms appearing in the decomposition (2.18) of the total variance in simple log-ratios. The first row of the table shows the sample clr-variances, which are also a decomposition of the total variance (2.19). The lower part of the Table 2.5 compares the sample centre of the composition with the percentages of overall Catalonia (Cat%), i.e. adding all votes across regions and then expressed as percentages. The Cat% row shows some important regional departures from the centre. Specially important is the deviation of C's, whose percentage in Cat% doubles that reported in the centre by regions. This is due to the high variance of this party across regions, as can be seen in the variation array (bold numbers). Also the variance of the log-ratio CiU over ERC is small (0.08) thus suggesting proportionality of these votes across the regions.

The singular value decomposition (SVD) is an important tool in statistics and, in particular, in CoDa analysis (Aitchison 1984). Its normal use is related to reduction of dimensionality, but it also provides an ilr transformation and the corresponding orthonormal basis. It starts with a matrix containing a log-ratio representation of the data set of size n, usually the clr-transformed data. Subtracting the mean value of the clr-components an (n, D)-matrix $\mathbf{M}$ is obtained. $\mathbf{M}$ is decomposed using SVD, $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$, where $(\cdot)^\top$ stands for transpose. The

Table 2.5 Variation array for Cat10. Upper triangle: simple log-ratio means; lower triangle: simple log-ratio variances. Total variance is 0.9952. The first row contains clr-variances. The centre of the regions is reported as cen%; the last row is the percentage computed from the total votes in Catalonia.

	abst	nota	null	C's	CiU	ERC	ICV	PSC	PP	other
clrVar	0.06	0.07	0.10	**0.22**	0.07	0.12	0.08	0.08	0.11	0.10
abst		0.067	0.15	0.29	0.07	0.19	0.08	0.05	0.10	0.14
nota	3.01		0.08	**0.53**	0.03	0.09	0.13	0.10	0.24	0.11
null	4.22	1.21		**0.77**	0.07	0.05	0.25	0.21	0.36	0.13
C's	3.76	0.75	−0.46		**0.55**	**0.85**	0.26	0.31	0.16	**0.59**
CiU	0.34	−2.67	−3.88	−3.42		0.08	0.15	0.14	0.26	0.07
ERC	1.95	−1.07	−2.28	−1.82	1.60		0.24	0.26	0.45	0.10
ICV	2.55	−0.46	−1.67	−1.25	2.21	0.61		0.10	0.18	0.17
PSC	1.40	−1.61	−2.82	−2.36	1.06	−0.54	−1.15		0.12	0.26
PP	2.02	−0.99	−2.20	−1.74	1.68	0.07	−0.54	0.62		0.38
other	1.75	−1.27	−2.48	−2.01	1.40	−0.20	−0.81	0.34	−0.27	
Cen%	38.9	1.9	0.57	0.91	27.6	5.6	3.0	9.6	5.2	6.8
Cat%	40.1	1.8	0.4	2.0	22.9	4.2	4.4	10.9	7.3	6.0

diagonal (D, D)-matrix $\mathbf{\Lambda}$ contains the singular values. Starting with clr-transformed data, the last of these singular values must be null. The squares of the singular values add up to the total variance and are variances of ilr-components (2.20). The $\mathbf{V}$ term is a (D, D)-matrix which reduces to a $(D, D-1)$-matrix after removal of the column corresponding to the null singular value. Its $D-1$ columns are the clr transformation of the elements of an orthonormal basis of the simplex. After removal of the column which corresponds to the null singular value, the rows of the $(n, D-1)$ matrix $\mathbf{U\Lambda}$ are the ilr coordinates of the centred data with respect to the orthonormal basis of the simplex defined by $\mathbf{V}$ (see Chapter 11). The biplot of this SVD (Gabriel 1971; Aitchison and Greenacre 2002) consists in representing a projection (usually bidimensional) of the first orthonormal components, called principal components, in the same plot as the projection of the centred clr variables. The centred clr-variables are scaled by the singular values so that the rays from the origin have length proportional to the standard deviation of the variables. The links between the vertices of rays are proportional to the standard deviation of the simple log-ratio of the parts corresponding to the rays. The cosine of the angle between two links approaches the correlation coefficient between the corresponding simple log-ratios; orthogonality of links suggests uncorrelation of the simple log-ratios. Figure 2.2 shows the CoDa-biplot for Cat10 explaining 83.5% of the total variance. The first principal component is clearly related to the links between C's and ERC, CiU, null, nota. This corresponds to the variances of simple log-ratios involving C's (Table 2.5). A political interpretation of this first principal component may be: positive values correspond to a larger proportion of non supporters of Catalan nationalism, while negative values correspond to a larger proportion of different versions of Catalan nationalism. The second principal component can also be interpreted as a kind of non-conservative parties or coalitions (positive values) against conservative parties (including the socialist PSC). Abstention is badly represented in the biplot and corresponds to a low variability principal

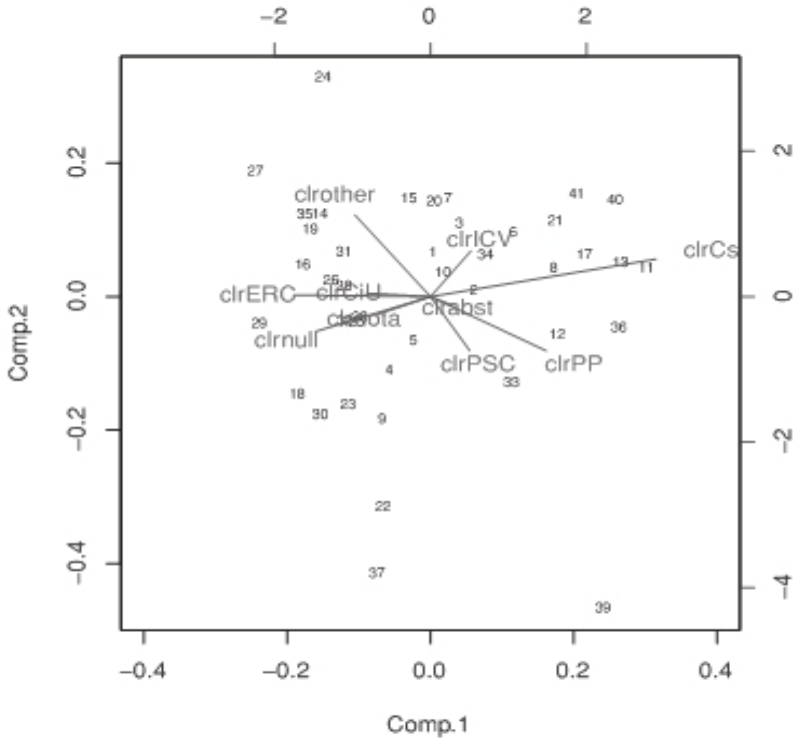

Figure 2.2 CoDa-biplot explaining 83.5% of total variance. Rays proportional to the standard deviation of clr variables. Regions: 1, Alt Camp; 13, Barcelonès; 41, Vallès Oriental; 39, Vall D'Aran (outlier).

component. Regions shown in Table 2.1 are numbered as 1 for Alt Camp, 13 for Barcelonès and 41 for Vallès Oriental. Vall D'Aran, treated as an outlier in the regression of Section 2.4, appears far apart from the centre in the lower part of the biplot (number 39).

Singular value decomposition and the CoDa-biplot determine an ilr transformation of the data, but the coordinates corresponding to principal axes are log-contrasts usually involving all parts with rather different coefficients. Interpretation of such principal components may be difficult and/or vague, although they correspond to the maximum variance explained. In many situations, the questions put forward by analysts suggest grouping parts of the composition and contrasting such groups. In those cases, a user designed basis of the simplex may be more understandable than the principal axes obtained using SVD. The technique is based on an SBP of the compositional vector.

The structure of the SBP, the ilr decomposition of the total variance (2.20), and the mean and dispersion of the sample coordinates can be summarized in a so called CoDa-dendrogram. Figure 2.3 shows the CoDa-dendrogram for Cat10 data following the SBP shown in Table 2.6. The vertical bars connect two groups of parts. The length of these bars is only that required to link the labels; but all of them, whatever their length, are equally scaled, in this case as a segment $(-4, +4)$. On each vertical bar a box-plot of quantiles (0.05, 0.25, 0.50, 0.75, 0.95) of the corresponding balance is represented to visualize sample dispersion. For instance the location of the box-plot of balance between C's and the rest of the parties shows that C's gets few votes in comparison with the other parties and coalitions. Anchored on each

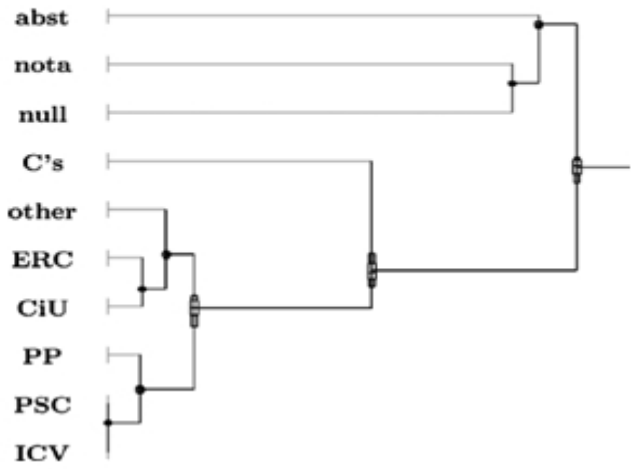

Figure 2.3 CoDa-dendrogram for Cat10 data. SBP as given in Table 2.6. Total variance is 0.9952. Scale of vertical bars ranges from −4 to +4.

vertical bar and pointing to the right, there is a horizontal bar. Its length is proportional to the variance of the balance, the ilr-variance in Equation (2.20). Contact of the vertical bar with the left end point of the horizontal bar is the mean balance (coordinate of the sample centre). Comparison of the mean balance and its median (0.50 quantile) in the box-plot gives an idea of the symmetry of a sample balance.

Figure 2.3 shows the relatively small variance of balances involving abstention, null and nota votes. The largest variance corresponds to the partition separating C's from the rest of the parties and coalitions (b_4), followed by b_5, which groups Catalan-only parties and coalitions (CiU, ERC, other) from lists of candidates with a presence in the whole of Spain (ICV, PSC, PP). Further partitions within these groups have smaller variances, thus indicating that the most variable balances across regions are precisely b_4 and b_5. Both balances are quite correlated (−0.835), thus showing that regions with a large relative presence of ICV, PSC, and/or PP in front of CiU, ERC and other are also those with a larger presence of C's.

Table 2.6 Sequential binary partition of Cat10, represented in Figure 2.3 as a CoDa-dendrogram.

	abst	nota	null	C's	CiU	ERC	ICV	PSC	PP	other
b_1	1	1	1	−1	−1	−1	−1	−1	−1	−1
b_2	1	−1	−1	0	0	0	0	0	0	0
b_3	0	1	−1	0	0	0	0	0	0	0
b_4	0	0	0	1	−1	−1	−1	−1	−1	−1
b_5	0	0	0	0	1	1	−1	−1	−1	1
b_6	0	0	0	0	−1	−1	0	0	0	1
b_7	0	0	0	0	−1	1	0	0	0	0
b_8	0	0	0	0	0	0	−1	−1	1	0
b_9	0	0	0	0	0	0	−1	1	0	0

2.6 Conclusion

Compositional data appear frequently in most experimental scientific fields. The statistical analysis of compositional data requires attention to their characteristics in order to avoid misleading results and conclusions. Spurious correlation is one of the most striking dangers when analysing compositional data; it was detected by K. Pearson at the end of the nineteenth century. The principles of compositional data analysis, established by John Aitchison in the 1980s, allow the development of a methodology free of all these problems. It relies on the study of log-ratios that transform the parts of a composition (e.g. concentrations in percentage, parts per one, parts per million) into real random variables that can be studied with conventional statistical methods. The developments related to the geometry of the simplex, called Aitchison geometry, allows the representation of the compositions in real orthogonal coordinates in such a way that usual statistical methods can be applied systematically. The use of coordinates in the simplex facilitates the computations and the analysis, but requires interpretation of (log)-ratios to obtain conclusions. That way, one will not be tempted to interpret each element individually, without acknowledgement of their relative character.

Acknowledgements

This research has been supported by the Spanish Ministry of Science and Innovation (projects CSD2006-00032 and MTM2009-13272) and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya (Ref. 2009SGR424).

References

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 1984 Reducing the dimensionality of compositional data sets. *Mathematical Geology* **16**(6), 617–636.

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

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

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

Aitchison J and Greenacre M 2002 Biplots for compositional data. *Applied Statistics* **51**(4), 375–392.

Aitchison J, Barceló-Vidal C, Egozcue JJ and Pawlowsky-Glahn V 2002 A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In *Proceedings of IAMG'02 – The VIII Annual Conference of the International Association for Mathematical Geology* (ed. Bayer U, Burger H and Skala W), vol. I and II. Selbstverlag der Alfred-Wegener-Stiftung, Berlin (Germany). pp. 387–392.

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

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

Buccianti A and Pawlowsky-Glahn V 2005 New perspectives on water chemistry and compositional data analysis. *Mathematical Geology* **37**(7), 703–727.

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

Connor RJ and Mosimann JE 1969 Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* **64**(325), 194–206.

Egozcue JJ 2009 Reply to 'On the Harker variation diagrams; . . .' by J. A. Cortés.. *Mathematical Geosciences* **41**(7), 829–834.

Egozcue JJ and Pawlowsky-Glahn V 2005a Coda-dendrogram: a new exploratory tool. In *Proceedings of CoDaWork'05, The 2nd Compositional Data Analysis Workshop* (ed. Mateu-Figueras G and Barceló-Vidal C). http://ima.udg.es/Activitats/CoDaWork05/. University of Girona, Girona (Spain).

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

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

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

Gabriel KR 1971 The biplot – graphic display of matrices with application to principal component analysis. *Biometrika* **58**(3), 453–467.

Martín-Fernández JA, Barceló-Vidal C and Pawlowsky-Glahn V 2003 Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology* **35**(3), 253–278.

Mosimann JE 1962 On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions. *Biometrika* **49**(1–2), 65–82.

Palarea-Albaladejo J, Martín-Fernández JA and Soto JA 2011 C-means clustering of compositional data. *Journal of Classification* (in press).

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

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

Pawlowsky-Glahn V and Egozcue JJ 2006 Análisis de datos composicionales con el coda-dendrograma. In *Actas del XXIX Congreso de la Sociedad de Estadística e Investigación Operativa (SEIO'06)* (ed. Sicilia-Rodríguez J, González-Martín C, González-Sierra MA and Alcaide D). Sociedad de Estadística e Investigación Operativa, Tenerife (Spain). CD-ROM.

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.

Thió-Henestrosa S, Egozcue JJ, Pawlowsky-Glahn V, Kovács LO and Kovács G 2008 Balance-dendrogram a new routine of codapack. *Computer and Geosciences* **34**(12), 1682–1696.