The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies

John Aitchison

Department of Statistics, University of Glasgow
Email: john.aitchison@btinternet.com

Abstract

In any discipline, where uncertainty and variability are present, it is important to have principles which are accepted as inviolate and which should therefore drive statistical modelling, statistical analysis of data and any inferences from such an analysis. Despite the fact that a single simple principle has existed over the last two decades and from this principle a sensible, meaningful methodology has been developed for the statistical analysis of compositional data, the application of inappropriate and/or meaningless methods persists in many areas of application. This paper identifies a number of common fallacies, confusions and misunderstandings in compositional data analysis with illustrative examples and provides readers with necessary, and hopefully sufficient, arguments to persuade the culprits why and how they should amend their ways. The paper is deliberately provocative in the hope that it may lead to constructive discussion.

Keywords: Basic principle of compositional data analysis; relation of pure mathematics to statistical modelling and analysis, scattergrams (Fenner, Harker, Pearce, ternary diagrams), scientific method (data snooping and testing hypotheses suggested by the data), transformation methodology, staying-in-the-simplex methodology.

0. Introduction: Some basic principles

0.1 The *basic principle of compositional data analysis.*

In any discipline where uncertainty and variability are present it is important to have principles which are accepted as inviolate and which should therefore drive statistical modelling, statistical analysis of data and any inferences from such an analysis. When in any discipline we say that a problem is compositional we are recognizing that the sizes of our specimens are irrelevant. For example, a geologist talking about the composition of an object, such as the major oxide composition of a rock, is admitting that interest is in a dimensionless problem. There is no concern about whether the rock specimen weighs one gm or one lb. Similarly in the study of the dietary content of cows' milk interest will focus on the dietary composition – proportions by weight of the total dietary content of the parts - protein, milk fat, carbohydrate, calcium, sodium, potassium – rather than on the size of the milk sample. Such trivial admissions have far-reaching consequences. Let us apply some clear thinking, as in Aitchison (1997), to acceptance of this fundamental scale invariance principle.

A simple example can illustrate the argument. In Fig.0.1, which shows 3-dimensional positive space R_+^3, the two points w [1.6, 2.4, 4.0] and W [3.0, 4.5, 7.5] represent the weights of the three parts [a, b, c] of two specimens of total weight 8 gm and 15 gm, respectively. If we are interested in compositional problems we recognize that these are of the same composition with the difference in weight being taken account of by the scale relationship $W = (15/8)w$. More generally two compositions *w* and *W* are compositionally equivalent, written $W \sim w$, when there exists a positive proportionality constant *p* such that $W = pw$. The fundamental requirement of compositional data analysis can then be stated as follows. Any meaningful construct or function *f* of a composition must be such that $f(W) = f(w)$ when $W \sim w$ or, equivalently,

$$f(pw) = f(w) \text{ for every } p>0. \tag{0.1}$$

This is a common problem in mathematics in group theory: the requirement (0.1) is that the function *f* must be invariant under the group of scale transformations. A general result of group theory is that any group invariant function can be expressed as a function of a maximal invariant. Now a function *h* is a maximal invariant when $h(W) = h(w)$ implies $W \sim w$. Here it is trivial to show that the (D-1)-vector function

$$h(w) = (w_1 / w_D, \ldots, w_{D-1} / w_D) \tag{0.2}$$

is a maximal invariant. The important consequence of this is the following.

Basic principle of compositional data analysis

> ***Any meaningful function of a composition can be expressed in terms of ratios of the components of the composition. Perhaps equally important is that any function of a composition not expressible in terms of ratios of the components is meaningless.***

Note that there are many equivalent sets of ratios which may be used for the purpose of creating meaningful functions of compositions. For example, a more symmetric set of ratios such as $w / g(w)$, where $g(w) = (w_1 \cdots w_D)^{1/D}$ is the geometric mean of the components of w, would equally meet the scale-invariant requirement.

All that this blinding by mathematics is saying is surely the obvious. Compositions are concerned with relative values and so ratios of components.

When I first became interested in compositional data I thought that this was self-evident, but apparently not. See, for example, the sequence of letters (Aitchison, 1989, 1990a, 1991a, b), Watson (1990, 1991)

in *Math Geology*, arising from Watson and Philip (1989) and ending with Aitchison (1992).

0.2 *The use of the subcompositional coherence argument*

In the past I have put forward subcompositional coherence as a second principle of compositional data analysis and have used it to emphasize the folly of using the Galton-Pearson correlation coefficient as a measure of dependence between raw components of a composition. My usual discussion is within the context of a practical situation.

> Consider two scientists A and B interested in soil samples, which have been divided into aliquots. For each aliquot A records a 4-part composition [animal, vegetable, mineral, water]; B first dries each aliquot without recording the water content and arrives at a 3-part composition [animal, vegetable, mineral]. Let us further assume for simplicity the ideal situation where the aliquots in each pair are identical and where the two scientists are accurate in their determinations. Then clearly B's 3-part composition $[s_1, s_2, s_3]$ for an aliquot will be a subcomposition of A's 4-part composition $[x_1, x_2, x_3, x_4]$ for the corresponding aliquot related by
>
> $$[s_1, s_2, s_3] = [x_1, x_2, x_3]/(x_1 + x_2 + x_3)\,. \qquad (0.3)$$
>
> It is then obvious that any compositional statements that A and B make about the common parts, animal, vegetable and mineral, should agree. This is the nature of subcompositional coherence.
>
> Consider the simple data set:
>
> | Full compositions $[x_1, x_2, x_3, x_4]$ | Subcompositions $[s_1, s_2, s_3]$ |
> |---|---|
> | [0.1, 0.2, 0.1, 0.6] | [0.25, 0.50, 0.25] |
> | [0.2, 0.1, 0.1, 0.6] | [0.50, 0.25, 0.25] |
> | [0.3, 0.3, 0.2, 0.2] | [0.375, 0.375, 0.25] |
>
> Scientist A would report the correlation between animal and vegetable as corr(x_1, x_2) = 0.5 whereas B would report corr(s_1, s_2) = -1. There is thus incoherence of the product-moment correlation between raw components as a measure of dependence.
>
> Note, however, that the ratio of two components remains unchanged when we move from full composition to subcomposition: $s_i / s_j = x_i / x_j$, so that, as long as we work with scale invariant functions, or equivalently express all our statements about compositions in terms of ratios, we shall be subcompositionally coherent.

My reason for downgrading subcompositional coherence as a principle to a convenient way of countering the use of product-moment correlation between raw components is that there are many statistical procedures where we would not expect analyses of full compositions and of subcompositions to produce relatable results, for example in any eigen-analysis such as principal component analysis. We have only to look at such analyses in R^D to see that principal component analysis of subvectors would have no direct relation to a principal component analysis of the complete vectors. So my advice here is to limit use of subcompositional coherence as an argument against the many surviving users of product moment correlations of raw compositional components.

This is probably the place to provide a caveat on the use of the symmetric centred log-ratio correlation matrix. It is tempting to think that

$$corr(\log(x_1 / g(x)) / \log(x_2 / g(x))\,, \qquad (0.4)$$

where $g(x)$ is the geometric mean $(x_1 \cdots x_D)^{1/D}$ of the components of the *D*-part composition *x*, is a reasonable way to measure the dependence between parts 1 and 2 of the composition. But this is subcompositionally incoherent, as any simple example will show. The reason for this incoherence is, of course, that use of a subcomposition automatically changes the dividing geometric mean.

0.3 *Basic principles of statistical analysis*

I often lie awake at night and worry about the state of much data analysis and the extent to which basic scientific principles are broken. In my consultative experience it has not been uncommon to be consulted after the experiment has been completed. The person consulting has collected data, filled the available filing space, does not know how to analyze the data and decides that the help of a statistician is required, with no clear idea of what specific questions are answerable by the experiment.

The basic scientific principle, which is so often ignored, is that, prior to experimentation or data collection, the specific questions, which the observations are expected to answer, should be clearly identified. I have no need to apologize for using yet again a statement (Jeffreys, 1961) by a great scientist, mathematician and statistician, Sir Harold Jeffreys, which should be a mantra of every investigator of uncertainty and variability:

> ***It is sometimes considered a paradox that the answer depends not only on the observations but on the question: it should be a platitude.***

In Cambridge when I studied for the Diploma in Mathematical Statistics my knuckles would have been rapped if I had dared to look at the data before I had identified relevant hypotheses to test to answer the experimenter's or observer's question. Changed days! It now seems to be common practice to data snoop, to spot possible connections and so formulate a hypothesis which is then tested on the snooped data, and of course the test found to be non-significant and so the hypothesis supported. Can anything be further from good scientific practice? And sadly there have been a number of examples in our CoDaWork workshops here in Girona. I'll look at one particular example later in this paper.

So I suggest that a protocol for any experimental or observational investigation might run along the following lines.

a) Before experimentation, specify all the questions of interest.
b) Design an experiment which is capable of answering these questions
c) Identify a sensible record set (sample space) for representation of the data?
d) Identify any concomitant information? And, if so, identify a sensible concomitant space?
e) If a parametric approach is feasible, identify a parametric class of (conditional) distributions that may be suitable.
f) Construct an appropriate lattice of hypotheses capable of answering all the questions.
g) Apply estimation and testing procedures to arrive at a working model.
h) If required, produce a predictive distribution for the working model.
i) Explain in as simple terms as possible inferences and conclusions to the experimenter.

Essentially this is a plea for good scientific practice and corresponding statistical design, modelling and analysis of experiments and observational studies.

1 Fallacies relating to 'traditional' methods of analysis

1.1 Fallacy *1a. Standard multivariate statistical analysis designed for real data works on compositional data*

The thinking behind this fallacy is, I suppose, that compositions are vectors of real numbers and the constraint simply confines the vector to a subspace embedded in real space, so real analysis must be suitable. The fallacy is not confined to compositional analysis, but has applied to early analysis of directional data and still is prevalent in the analysis of positive vectors, again since $R_+^D \subset R^D$. I have

asked many fellow statisticians, particularly those involved in consultative work, if they have ever had multivariate consultative work involving vectors with negative components and the answer, on the whole, is a resounding 'No'; and, when further asked whether they can envisage such data, the responses usually involve situations with arbitrary origins. The fact that the properties of real space are so familiar, even to school children, has made it easy to produce a statistical methodology relevant to real vectors, with the unfortunate consequence that any vector data that contains real numbers are flung into the R^D mill. It is, of course, now well known that positive multivariate data can be treated effectively by a logarithmic transformation, followed by standard R^D analysis or, equivalently, within R_+^D, with appropriate operations of change (proportional change), of powering and the obvious logarithmic metric.

Historically it was Chayes (1960, 1962, 1971) who first persistently pointed out to geologists that the application of the then available multivariate statistical tools, designed for analysis of data vectors in *D*-dimensional real space R^D, to compositional data was not only flawed but meaningless. Chayes tackled, in particular, the impossibility of interpreting the product moment correlation coefficient between two parts of a composition. Since the whole structure of standard multivariate statistical analysis depends on these concepts of product-moment covariance and correlation, it should have been obvious that standard multivariate statistical analysis should be placed in a compositional garbage bin. Chayes and colleagues attempted to study the pathology of the application of these standard methods to compositional data, presumably in the hope of finding adjustments which might alleviate the situation. Unfortunately, as Chayes and his colleagues soon found out, pathology does not cure a patient. Appropriate treatment is required to ease his or her disease.

To an extent at CoDaWork'08 this may seem to be preaching to the converted. But I think it is worth pointing out that this fallacy is still prevalent. I can assure you that a tour of the web under 'compositional data analysis' will find many examples. What advice can be provided on how to counter such applications of inappropriate methodology? My first approach is always to point out the basic principle of compositional data analysis with illustrations. Probably the more powerful argument there is to demonstrate, as in Section 0.2, subcompositional incoherence in the use of product moment correlation as a means of conveying a measure of dependence between two parts.

A failure here may mean that you are not dealing with someone who can claim to be a scientist, though such a statement will get you nowhere. One of the difficulties, I think, is that many, probably most, scientists in their brush with statistics have been inculcated with the idea that correlation is the be-all-and-end-all of the statistical measure of dependence. I have come across this is many forms. Unfortunately it lingers with many referees of papers on compositional data. Let me relate a number of instances.

Many recent fallacies are subtle restatements of Fallacy 1a in convoluted language. They often typically arise in the refereeing of papers using the principle outlined above. In this refereeing there is commonly reference to knowledge derived from *traditional methods,* which can only mean the use of standard R^D multivariate statistics. One such fallacy is the following.

1.2 Fallacy *1b. The fallacy of hanging on to the concept of correlation in considering two parts of a composition*

Here are examples of such restatements.

Subscribers to this fallacy clearly are attempting to defend themselves behind the bulwark of tradition. Inferences from past data analysis, however arrived at, are sacrosanct despite the fact that pre-1980 they could have been based only on standard multivariate statistical analysis.

I use my own experience of attempts to combat such refereeing experiences to demonstrate the nature of this type of fallacy.

In one paper I attempted to explain that when dealing with a typical *D*-part composition $[x_1,\ldots,x_D]$, the nearest useful characteristic to the meaningless $corr(x_i, x_j)$ is the characteristic $\tau_{ij} = \mathrm{var}\{\log\{x_i / x_j\}\}$, what I had called the relative variance. I had pointed out that a zero value of τ_{ij} means that the ratio x_i / x_j is constant; in other words, component *i* is a fixed proportion of component *j*, surely the strongest form of dependence between the two parts when we recall from the basic principle that any meaningful function of two parts must involve this ratio. The practitioner of Fallacy 1b may say: OK, but you have not given me some means of saying when parts *i* and *j* are uncorrelated. The answer here is, of course, that the request is coming from someone who persists in the use of product-moment correlation. The only useful reply to this fallacy is to point out that when attention is restricted to two parts of a composition the larger the value of τ_{ij} the more 'uncorrelated/independent' they will appear to a traditionalist.

Here is another comment from a referee.

> There is a major discrepancy between logratio modelling and the intuitive approach of the petrologist/geochemist. The authors state that 'there is no hope of the product-moment correlation serving as a usable substitute for the relative variation'. This is a very strong statement which is probably not correct because two variables with a relatively low relative variation could be strongly correlated in the product-moment sense.
>
> The following hypothetical example of a major oxide (*x*) and a trace element (y) illustrates this:

x (in weight %)	*y* (in ppm)	log(*x*/*y*)
2	1	0.69315
20	20	0
21	21	0
18	18	0

> The product moment correlation coefficient is $corr(x, y) = 0.99997$. The variance, unbiased estimate of the logratio (base e), is 0.120113. Consequently excellent 'correlation' between *x* and y is suggested by corr(*x*,*y*), but relatively poor correlation by the relative variation. This example is, of course, artificial. However a situation of this type could arise if the trace element (y) would be chemically associated only with a single major-oxide (x).

It is clear from my remarks above that the referee has missed the whole point of the relative variance, in that he expects a large value of relative variance to characterise some form of high dependence whereas the opposite is the case.

My response to the referee is possibly worth recording.

> You seem to be suggesting that a low relative variation would be in conflict with a highly positive crude correlation; the contrary is the case. The lower the relative variation the more proportional the two components are and so, in so far as it is possible to place an interpretation on crude correlation, the lower the relative variation the higher the positive crude correlation. Your example therefore is not disturbing: a low relative variation and a high crude correlation. It does, however, suggest to me another avenue of persuasion that relative variation has got 'everything' and crude correlation 'nothing'. Consider the following three data sets of two-part compositions and the corresponding relative variations and crude correlations.

1		2		3	
x_1	x_2	x_1	x_2	x_1	x_2
28	72	20	80	10	90
29	71	25	75	20	80

	30	70	30	70	30	70
	31	69	35	65	40	60
	32	68	40	60	50	50
$corr(x_1, x_2)$	-1		-1		-1	
relvar	0.0845		0.150		0.738	

Which of the measures contains useful information? Answer: the relative variance, because it works with ratios and the range of ratios is very different for each sample, for (1) 0.39 – 0.47; for (2) 0.25 – 0.67, for (3) 0.11 – 1.00.

1.3 Fallacy *1c. The new methodology merely confirms what we already know about this situation through the use of traditional methods*

The only comment here is probably to say: Well, either you were lucky with your traditional methods, or at least the new methodology must be correct in this case.

1.4 Fallacy *1d. Any new methodology must surely support the results of long standing traditional analysis*

This, I have found, is a great weapon of destructive criticism by referees. If the new methodology does not support traditional views of a situation, then the only recourse is to question how the traditional view was arrived at. If previous data are available, analyse the data and go on from there.

1.5 Fallacy *1e. Knowledge of the discipline can overcome the constant-sum constraint and allow the use of* R^D *multivariate methodology*

In my experience, consultative statisticians spend much time with their consultee – doctor, economist, geologist, biologist, whomever – in an attempt to make sure that the statistical modelling, analysis and inferences are appropriate to the nature of the discipline. Not so with two papers by Woronow (1997 a, b) stating the above fallacy (or conceit), with a cavalier dismissal of all the attempts by statisticians over a period of fifteen years to help the discipline of geology. There is little point in reviewing the many basic errors in these papers since they have been thoroughly dealt with in Aitchison (1999).

We turn now to a number of graphical fallacies.

2 Fallacies relating to graphical representations

2.1 Fallacy 2a. That scattergrams of two raw components provide useful information about compositional variability

Despite its sentence of death by Chayes and others in the 1960s, the Harker diagram (Harker, 1909), now almost a hundred years old, is still alive, though hardly well. A Google expedition under the entry "Harker diagram" at this time of writing (27 February 2008) found 623 sites. We might hope, even expect, that many of these sites would be continuing the Chayes exposure of the deceptiveness of Harker diagram use. Alas, this is not so, and many of the sites even consist of the instruction of geology students on how valuable Harker diagrams are in understanding geochemistry, a sad reflection on geological instruction in some institutions. The Harker diagram in its geological version is essentially a scattergram of two raw components of compositions with SiO_2 as the horizontal variable. Other scattergrams of two components, such as the Fenner diagram (Fenner, 1913) are also still in use. The Chayes condemnation was essentially based on the uninterpretable nature of the Galton-Pearson correlation coefficient of raw components and there is little need here to reiterate the Chayes argument.

We shall take a recent misuse of Harker diagrams to illustrate this persistent fallacy, and to attempt to provide some ideas of how to combat the fallacious argument and provide a sensible analysis of the underlying compositional problems. At CoDaWork'03 Baxter et al (2003) considered problems in the compositional data analysis of 63 colourless Roman-British glass facet-cut beakers. In their analysis they presented in particular a scattergram of (Al, Fe) as in Fig. 2.1. In their interpretation of this scattergram they commented that 'it can be seen that there is a distinct compositional group of low Al and relatively high Fe facet-cut beakers, labelled (as in Fig. 2.1) with 'o' in the plot, that is also compositionally distinct from other types labelled x in Fig. 2.1. But, if we pay attention to the basic principle of compositional data analysis, the scattergram allows only the investigation of the ratios Fe/Al. Fig.2.2 allows this investigation by showing the origin of the Al, Fe scales and the ratios Fe/Al in the slopes of the lines from the origin to the compositional points. The situation is very different and there is less convincing separation. Baxter et al (2003) go on to create raw component and logratio biplots. They claim that inspection of their raw component biplot comfortably suggests their previously noted grouping from the Al-Fe scattergram. This is not surprising since their division of the 63 compositional points is made on a linear basis and the raw component biplot is based on linear mathematics. Their logratio biplot also shows separation; this would follow from the fact that there is separation of the ratios in Fig.2.2. It is difficult to assess their statement that 'separation is not as clearly seen as in the raw component biplot', based on subjective visual judgment. The criticisms of these analysis are threefold.

First and probably most important scientifically, is that the authors seem to have fallen into the trap of data-snooping, examining the data before they have framed a question or hypothesis about the situation. If the data suggest a hypothesis it is not uncommon for that hypothesis to be 'accepted' by analysis of the data which suggested the hypothesis. In particular if such a compositional hypothesis is suggested from the use of linear ideas applied to the data then it is not surprising that subsequent linear analysis will support the hypothesis.

The construction of biplots for all 63 compositions does not address the 'diagnostic' problem of how separate the two 'groups' are and to what extent the separation may depend only on some subcompositional information. Of course, in the present case this is not a straightforward analysis. The suspect identifying of groups as in Figs 2.1, 2.2 above means that there is technically what is known as complete separation and so, for example, a binary logistic approach to attempt to identify important subcompositions along the lines of Thomas and Aitchison (1998) is not available, although the methodology of dealing with the complete separation problem in Aitchison, Kay and Lauder (2004) could be used.

2.2 Fallacy 2b. The ternary diagram is better than the Harker diagram

I think the way in which the ternary diagram is generally used is no better than Harker or Fenner diagrams. The reason for this is that the components of a 3-part composition are thought of as the perpendicular distances from the representational point to the sides of the triangle. This is no more than viewing the ternary diagram as a scattergram of the raw components similar to the Harker diagram's treatment of two raw components.

Let me illustrate what I mean by examining a particular ternary diagram. Table 1 records 40 3-part compositions with parts [1, 2, 3]. The corresponding ternary diagram is shown in Fig 2.3. In the diagram I have shown what might be considered an obvious separation of the points into groups, 13 identified by an o, those with A-component greater than 0.65 and 27 identified by an x, with A-component less than or equal to 0.65. Again we emphasize that any subsequent investigation of any hypothesis about this separation based on the snooped data is suspect. I need not give details except to say that the results would be similar to those for the British-Roman beakers above.

The spurious nature of all this analysis is evident when I reveal that the 40 3-part compositions of Table 1 arose from a simulation of 40 3-part compositions from a $L^3(\xi, \mathrm{T})$ distribution with $\xi = [0.5,\ 0.3,\ 0.2]$ and

$$\mathrm{T} = \begin{bmatrix} 0.00 & 2.00 & 1.20 \\ 2.00 & 0.00 & 0.50 \\ 1.20 & 0.50 & 0.00 \end{bmatrix}.$$

My biggest mistake when I became interested in compositional data and recognised that interest could only be in relative values was to be fixed on the simplex and its simple visual 3-part ternary diagram. What I should have done was to point out that a D-part composition should be recorded as a (*D*-1) vector of ratios in R_+^{D-1} .

2.3 What *are sensible compositional diagrams?*

The answer is easy: ratio scattergrams, Pearce diagrams, logratio scattergrams, and for a complete compositional picture, compositional biplots (Aitchison, 1990b; Aitchison and Greenacre, 2002) and conditional compositional biplots. Notice I have not included the recent coda dendrogram (Egozcue and Pawlowsky-Glahn, 2005) in this list. I will explain why later in this paper.

2.4 *One general word of caution about graphical approaches.*

In danger of being repetitive I make the following remarks about graphical approaches. When I was a statistics student in the 1950s great emphasis was placed on scientific method. Put simply, your hypothesis preceded your experiment. The purpose of your experiment was to produce data whereby you could test your hypothesis. If the hypothesis was rejected you thought a bit more deeply; and hopefully were led to another hypothesis to test. If not rejected, you had a working hypothesis on which to build a more detailed theory, hopefully giving rise to further hypotheses to test. I have a feeling that there has been considerable slippage from this view of scientific method. After the experiment data snooping, including graphical methods, suggests hypotheses which are subsequently tested. To me that seems poor science.

3 Fallacies involving ideas of unclosing or opening compositional data sets

There have been various approaches to the statistical analysis of compositions by attempts to construct 'open data' from which the closed sets may have arisen; and to infer some properties of the closed set by testing various hypotheses concerning the open set. Fortunately these, such as the early Chayes-Kruskal (1966) approach, have now gone out of fashion, though it is worth pointing out a recent attempt by Whitten (1995) to open up major oxide compositional data in proportions by weight by converting the weights to volume by use of specific gravities. Let $[x_1,\ldots,x_D]$ be a typical composition of *D* major oxides in weight proportions and let $[s_1,\ldots,s_D]$ be the specific gravities of the *D* major oxides. Then the conversion to volumes produces the 'open' set $v=[v_1,\ldots,v_D]=[x_1/s_1,\ldots,x_D/s_D]$. The concept of the open set is illusory, since the unit-sum constraint on *x* is simply replaced by an equally awkward constraint on *v*, namely $s_1v_1+\cdots+s_Dv_D=1$.

4. Some necessary definitions, concepts and notation

4.1 Introduction

It is convenient here to set out definitions, concepts and notation which will be used in discussion of two main directions in which compositional data analysis has developed in the last three decades – transformation methodology and a staying-in-the-simplex methodology. These methodologies, simply and properly applied, lead to equivalent inferences about any specific compositional question, and the choice of how to express and understand inferences seems to me to depend on how mathematically equipped the interpreter is. Despite this equivalence there have been some puzzling attempts to identify differences, which purport to suggest that the staying-in-the-simplex is somewhat superior. Later in this

paper I will attempt to specify where and why this staying-in-the-simplex approach has gone statistically wrong.

4.2 A little history of transformation methodology

My original, largely intuitive, approach to compositional data analysis in Aitchison (1981, 1982), Aitchison and Shen (1980), and in its extended form in Aitchison (1986a, b) was by way of a logratio transformation methodology. The rationale ran along the following lines: compositional vectors contain information only about relative values of their components, so first 'think ratios'. Then, on realising that quotients are more difficult to handle than sums or differences, 'think logratios'. The simplest logratio transformation seemed to be the additive logratio transformation $alr : S^D \to R^{D-1}$ defined by

$$z = alr(x) = [\log(x_1 / x_D) \ldots, \log(x_{D-1} / x_D)], \tag{4.1}$$

with inverse transformation $alr^{-1} : R^{D-1} \to S^D$ defined by

$$alr^{-1}(z) = C[e^{z_1}, \ldots, e^{z_{D-1}}, 1], \tag{4.2}$$

where C is the familiar closure-to-unity operation.

It is a one-to-one transformation between the unit simplex S^D and R^{D-1}, and so any questions concerning compositions are transferable to questions about the transformed compositions, and inference from analysis of the transformed data in real space by the use of standard multivariate analysis should be easily interpretable in terms of inverse transformations. This is a convenient place to define another useful transformation, the centred logratio transformation $clr: S^D \to U^{D-1}$, where U^{D-1} is the hyperplane $\{u \in R^D : u_1 + \ldots + u_D = 0\}$ in R^D, defined as follows:

$$u = clr(x) = [\log\{x_1 / g(x)\}, \ldots, \log\{x_D / g(x)\}] \quad (x \in S^D, u \in U^{D-1}), \tag{4.3}$$

where $g(x) = (x_1 x_2 \cdots x_D)^{1/D}$ is the geometric mean of the components of x. It has the advantage of treating the parts symmetrically.

Transformation techniques have been very popular and successful for more than a century, from the Galton-McAllister introduction of such an idea in 1879, in their logarithmic transformation for positive data, through variance-stabilising transformations for sound analysis of variance, to the general Box-Cox transformation and the implied transformations in generalised linear modelling. Nevertheless, they have had an interesting history and have not been without opposition. No less a scientist-statistician than Karl Pearson clearly misunderstood the nature of the transformation methodology.

> Supposing, as in English female crania, nasal breadth is asymmetrical, what is the quantity which is symmetrically distributed of which nasal breadth is a function? It has no reality in the organism at all.

And yet he seemed happy to use his 'frequency curves for all circumstances' uncritically to describe similar types of variability.

The logratio methodology, however, also drew fierce opposition from other disciplines, in particular from sections of the geological community. The opposition seemed to run roughly along the lines of the Pearson argument: what on Earth has log(MgO/CaO) to do with the geology of major oxides of rocks. The reader who is interested in following the arguments that have arisen may examine the letters to the Editor of Mathematical Geology over the period 1988 through 2002; in particular, see Watson and Philip (1989), Aitchison (1989, 1990a), Watson (1990), Aitchison (1991a), Watson (1991), Aitchison (1991b, 1992b), Woronow (1997a, 1997b), Aitchison (1999), Zier and Rehder (1998), Aitchison et al. (2000), Rehder and Zier (2001), Aitchison et al. (2001, 2002).

Statisticians also seem to have difficulties with logratio transformation methodologies. An anonymous referee of Aitchison (1983), which introduced compositional logcontrasts as the means of providing the equivalent of R^D principal component analysis for compositions, dismissed the logcontrast idea as unnecessary since R^D principal component analysis would immediately recognize and eliminate the constant-sum constraint corresponding to the zero eigenvalue and the other principal components would reveal the nature of the compositional variability. Also there is confusion in Gower (1987), in which the claim is made that the logarithmic transformation will not straighten the data of a data set associated with the famous Hardy-Weinberg law. The confusion is that it is logratio, not logarithmic, transformations that are relevant to compositional data analysis, and any compositional data analyst applying the *alr* transformation to that data set would see linearity and arrive at the Hardy-Weinberg law from the data analysis.

In many ways the adverse responses have helped to clarify the important principle underlying compositional data analysis and to consolidate knowledge of the underlying algebraic-geometric structure of the simplex sample space, the subject of the next section.

4.3 An algebraic-geometric structure on the simplex

In the discussion of Aitchison (1982) Dr N. I. Fisher, in drawing analogies with modelling problems involving directional data, suggested that greater insights might be obtained into the interpretation of compositional data analysis if concepts and analysis could be confined to the simplex rather than a process of transformation and inverse transformation. Understanding of any algebraic-geometric structure of a sample space is certainly a potential help in modelling practical statistical problems and although, in the case of the simplex sample space, the necessary ingredients appeared in Aitchison (1986), understanding has only gradually developed so that one structure is now that of a Hilbert space. Thus the way is now open for a full discussion of the relative merits, both from theoretical and practical viewpoints of two methodologies – transformation or staying-in-the-simplex.

Discussion in later sections will be simpler if we record here a pure mathematical structure which can be placed on the unit simplex choice as a compositional sample space.

The unit simplex as a vector space

Fundamental operations of change in the simplex are those of perturbation and power motivated and spelt out by Aitchison (1986, pp. 42 and 120). In their simplest forms these can be defined as follows. Given any two *D*-part compositions $x, y \in S^D$ their perturbation is

$$x \oplus y = C[x_1 y_1, \ldots, x_D y_D] \tag{4.4}$$

Clearly, the operation $\oplus$ defines an Abelian group, with identity element $e = (1/D)[1, \ldots, 1]$ and inverse

$$x^{-1} = C[1/x_1, \ldots, 1/x_D]. \tag{4.5}$$

An important concept here is how perturbation is related to change in a composition *x* to a composition *y*. The perturbation *p* which changes *x* to $y = p \oplus x$ can be written as

$$p = y \ominus x = C[y_1/x_1, \ldots, y_D/x_D]. \tag{4.6}$$

Given a D-part composition $x \in S^D$ and a real number $a \in R^1$ the power transformed composition is

$$a \otimes x = C[x_1^a, \ldots, x_D^a]. \tag{4.7}$$

Note that we have used the operator symbols $\oplus$ and $\otimes$ to emphasize the analogy with the operations

of translation and scalar multiplication of vectors in real space. It is trivial to establish that these operations define a vector or linear space structure on S^D.

Powering, as described above, may seem an esoteric operation, but it has importance for a number of reasons. First it may be of relevance directly because of the nature of the sampling process. For example, in grain size studies of sediments, sediment samples may be successively sieved through meshes of decreasing diameters and the weights of these successive separations converted into compositions based on proportions by weight. Thus, though separation is based on the linear measurement diameter, the composition is based essentially on a weight, or equivalently a volume measurement, with a power transformation being the natural connecting concept. Later in Sections 5.1 and 6.2 we shall see that the powering operation plays a central role in techniques of singular value decompositions of compositional data sets and can be useful in describing regression relations for compositions.

The unit simplex as a metric vector space

Aitchison (1983, 1986, p. 193) defined a simplicial metric principally to provide a duality between total variability of a compositional data set as measured by the trace of an estimated covariance matrix and the sum of mutual distances between compositions of the data set. The definition of the metric $\Delta_S(x,y)$ as a distance between two compositions $x, y \in S^D$ has many equivalent forms. These can be expressed in terms of different logratio transformations of the composition to real spaces, in particular the centred clr and additive *alr* logratio transformations. In terms of these transformations the definitions are

$$\Delta_S(x,y) = \left[\sum_{i=1}^{D} \left\{ \log \frac{x_i}{g(x)} - \log \frac{y_i}{g(y)} \right\}^2 \right]^{1/2} = [\{clr(x) - clr(y)\}\{clr(x) - clr(y)\}^T]^{1/2}, \tag{4.8}$$

$$\Delta_S(x,y) = [\{alr(x) - alr(y)\} H^{-1} \{alr(x) - alr(y)\}^T]^{1/2}, \tag{4.9}$$

where $H = [h_{ij}]$, with $h_{ij} = 2 \ (i = j)$, $h_{ij} = 1 \ (i \neq j)$ (Aitchison (1986, Section 4.7). Yet another equivalent definition is

$$\Delta_S(x,y) = \left[\frac{1}{D} \sum_{i<j} \left\{ \log \frac{x_i}{x_j} - \log \frac{y_i}{y_j} \right\}^2 \right]^{1/2}. \tag{4.10}$$

It is trivial to demonstrate that Δ_S satisfies all the requirements of a metric, namely

M1. Positivity: $\Delta_S(x,y) > 0 \quad (x \neq y), \quad = 0 \quad (x = y)$.

M2. Symmetry: $\Delta_S(x,y) = \Delta_S(y,x)$.

M3. Power relationship: $\Delta_S(a \otimes x, a \otimes y) = |a| \Delta_S(x,y)$.

M4. Triangular inequality: $\Delta_S(x,z) + \Delta_S(z,y) \geq \Delta_S(x,y)$.

Note that the power transformation property M3 is the analogue of the scalar multiple property of Euclidean distance in R^d.

The fact that this metric has also desirable properties relevant and logically necessary, such as scale, permutation and perturbation invariance, and subcompositional dominance, for meaningful statistical analysis of compositional data, is now well established and the relevant properties are recorded briefly here, for future discussion.

M5. Permutation invariance: $\Delta_S(xP, yP) = \Delta_S(x,y)$, for any permutation matrix P.

M6. Perturbation invariance: $\Delta_S(x \oplus p, y \oplus p) = \Delta_S(x, y)$, where *p* is any perturbation.

M7. Subcompositional dominance: If s_x and s_y are similar, say (1,...,C)-subcomposi-tions of x and y, then $\Delta_{S^C}(s_x, s_y) \le \Delta_S(x, y)$.

It is worth noting here that M7, subcompositional dominance, the equivalent of the obvious R^D metric property that the Euclidean distance between two points cannot be less than the Euclidean distance between the projections of these points on some hyperplane, was crucial in dismissing the Watson and Philip (1989) persistent claim that the *unique* measure of difference between two compositions was the *angle* between their ray representations.

The unit simplex as a Hilbert space

The extension of the simplex structure from metric vector space to Hilbert space came almost simultaneously with the recognition of the definition of the inner product of two compositions $x, y \in S^D$ as

$$\langle x, y \rangle = \sum_{i=1}^{D} \log \frac{x_i}{g(x)} \log \frac{y_i}{g(y)} \tag{4.11}$$

and the associated norm as

$$\| x \| = \left(\sum_{i=1}^{D} \left(\log \frac{x_i}{g(x)} \right)^2 \right)^{1/2} . \tag{4.12}$$

by Aitchison (2001), Billheimer, Guttorp and Fagan (2001), Pawlowsky-Glahn and Egozcue (2001). An interesting aspect of this extension is that an inner product $\langle b, x \rangle$ can be expressed as

$$\sum_{i=1}^{D} \log \frac{b_i}{g(b)} \log \frac{x_i}{g(x)} = \sum_{i=1}^{D} a_i \log x_i \tag{4.13}$$

where $a = \log\{b / g(b)\}$ and so $a_1 + \ldots + a_D = 0$. Thus, inner products play the role of logcontrasts, well established as the compositional 'linear combinations' required in many forms of compositional data analysis such as principal component analysis, (Aitchison, 1983) and the investigation of subcompositions as concomitant or explanatory vectors (Aitchison and Bacon-Shone, 1984; Aitchison, 1986, Chapters 8 and 12).

It is convenient here to add simple results on calculus in the simplex – differentiation, rates of change and integration. Clearly, in compositional processes rates of change of compositions are important and here we define the basic ideas. Suppose that a composition $x(t)$ depends on some continuous variable *t* such as time or depth. Then the rate of change of the composition with respect to *t* can be defined as the limit

$$Dx(t) = \lim_{dt \to 0} \frac{1}{dt} \otimes (x(t + dt) \ominus \mathrm{x(t)}) = \mathrm{C}(\exp(\frac{\mathrm{d}}{dt} \log x(t))), \tag{4.14}$$

where *d/dt* denotes 'ordinary' differentiation with respect to *t*. Thus, for example, if $x(t) = \xi \oplus h(t) \otimes \beta$, then $Dx(t) = h'(t) \otimes \beta$.There are obvious extensions through partial differentiation to compositional functions of more than one variable. We note also that the inverse operation of integration of a compositional function $x(t)$ over an interval (T_0, T) can be expressed as

$$\oint x(t)dt = C(\exp(\int_{T_0}^{T} \log x(t)dt)), \tag{4.15}$$

In what follows I shall use an application which seems popular with compositional data analysts, particularly in discussing possible doubts and difficulties of the transformation methodology. This is the Arctic lake sediment problem discussed by Aitchison (1986, Sections 7.6-7.8 and Data 5) to illustrate how to relate compositions to concomitant data through compositional regression.

5. Confusions and fallacies relating to transformation methodology

5.1 Confusions, fallacies and misunderstandings about *alr* transformation methodology: a reassessment of the *alr* transformation methodology

The *alr* transformation methodology has, in my view, withstood all attacks on its validity as a statistical modelling tool. Indeed it is an approach to practical compositional data analysis which I recommend particularly for non-mathematicians. The advantage of its logratios involving only two parts, in contrast to *clr* and *ilr* (isometric transformations discussed later in this section), which use logratios involving more than two and often many parts, makes for simple interpretation and far outweighs any criticism, more imagined than real, that the transformation is not isometric (I shall expand on this comment in Section 7). We first collect briefly a number of established results to allow discussion of the practicalities of the transformation.

Measure of 'central tendency': By analogy with definitions of centres of distributions associated with other sample spaces a least squares approach would identify $\xi = cen(x)$ as the $\xi \in S^D$ which minimises $\mathrm{E}(\Delta_S^2(x,\xi))$ and this can be expressed in a variety of ways as

$$cen(x) = C[\exp\{\mathrm{E}(clr(x))\}] = alr^{-1}\mathrm{E}(alr(x)) = C\{\exp(\mathrm{E}(\log \mathrm{x})\}. \qquad (5.1)$$

At first sight (5.1) seems a very unfamiliar object, until we realize that for any positive random vector $z \in R_+^D$ the formal definition of the *geometric* mean is $\exp\{\mathrm{E}(\log z)\}$, the vector of marginal medians. Note here that, although it seems that we have abandoned in the use of log(x) our scale-invariant principle of using only ratios, the complete expression for $cen(x)$ involves a closure operation which ensures ratios.

It is worth digressing here to demonstrate the practical implications of this simple result. For a typical compositional data set

$$X = \begin{bmatrix} \mathrm{x}_{11}\ \mathrm{x}_{12}\ \ldots\ \mathrm{x}_{1D} \\ \mathrm{x}_{21}\ \mathrm{x}_{22}\ \ldots\ \mathrm{x}_{2D} \\ \cdot\ \cdot\ \cdot\ \cdot \\ .\mathrm{x}_{N1}\ \mathrm{x}_{N2}\ \ldots\ \mathrm{x}_{ND} \end{bmatrix}, \qquad (5.2)$$

standard practice seems to be to take the arithmetic centre $\bar{x} = [x_{.1},,\ldots,x_{.D}]$, where $x_{.i} = \sum_{n=`}^{N} x_{ni} / N$.

What is being advocated here is the use of

$$C[g_1,\ldots,g_D] \qquad (5.3)$$

as centre of the compositional data set X, where

$$g_i = \left(\prod_{n=1}^{N} x_{ni}\right)^{1/N} \qquad (5.4)$$

is the geometric mean of the *i*th component over all *N* cases.

And there can be a substantial difference in the location of these centres. For example, Fig. 5.1 shows the arithmetic and geometric centres, A and G, for the 37 [sand, silt, clay] compositions of the Arctic lake sediment problem. It is clear that G is well within the distribution of compositions, whereas A appears to be an atypical composition.

A question still frequently asked about the *alr* transformation is: Do we get different results if we use a divisor different from x_D in the formation of logratios. The answer is no: all the statistical procedures are invariant under a permutation of the compositional parts. To a large extent I dealt with this question in Aitchison (1986, Chapter 7), but there remain doubts in some quarters about *alr* applications particularly in compositional regression applications. I could deal with these points theoretically, but I feel it is almost more convincing to illustrate the veracity of the equivalence of approaches having in mind a concrete example to illustrate the veracity of the invariance assertion.

Example. The relation of Arctic lake sediments compositions [*sand, silt, clay*] *to lake depth.* This was used as an illustrative example by Aitchison (1986, Sections 7.6 – 7.8). We first provide practical evidence of the claim of permutation invariance. In that analysis it was found through consideration of a lattice of possible hypotheses that a reasonable working model would be the compositional regression form

$$alr(x) = [\log(sand/clay), \log(silt/clay)] = [\alpha_1 + \beta_1 \log(depth), \alpha_2 + \beta_2 \log(depth)] + error\,,$$

arriving, after standard multivariate regression analysis and omitting the error terms,

$$\log(sand/clay) = 9.70 - 2.74\log(depth), \quad \log(silt/clay) = 4.80 - 1.10\log(depth)\,.$$

From this we can derive

$$\log(sand/silt) = 4.90 - 1.64\log(depth)\,.$$

If instead of the order [*sand, silt, clay*] we permute the parts to [*sand, clay, silt*] the regression analysis produces

$$\log(sand/silt) = 4.90 - 1.64\log(depth), \quad \log(clay/silt) = -4.80 + 1.10\log(depth)\,,$$

in conformity with the earlier permutation result. Moreover the residuals transformed in each case back to the [*sand, silt, clay*] ternary diagram by alr^{-1} are identical as shown in Fig. 5.2.

Comparison with the use of a *clr* regression of the form

$$clr(x) = [\gamma_1 + \delta_1 \log(depth), \gamma_2 + \delta_2 \log(depth), \gamma_3 + \delta_3 \log(depth)\,] + \text{error}$$

provides after estimation and omitting the error terms, the regressor function vector

[4.86 – 1.46 log(*depth*), -0.03 +0.18 log(*depth*), -4.83 + 1.28 log(*depth*)],

again in conformity with the *alr* approach, for example giving the sand/clay relationship by subtraction of the third from the first regressor function as 9.80 - 2.74 log(*depth*). Again the [*sand, silt, clay*] residuals are identical with those in Fig. 5.2.

We compare these results with the use of an *ilr* regression. One of the simplest *ilr* transformations here is

$$ilr(x) = [(\log(sand) - \log(silt))/\sqrt{2},\ (\log(sand) + \log(silt) - 2\log(clay))/\sqrt{6}\,]\,.$$

Regressing the transformed data against log(depth) produces a regressor function vector

$$[3.46 - 1.16 \log(depth),\ 5.92 - 1.57 \log(depth)].$$

To return to the result for the alr transformation applied to [*sand, silt, clay*] we have to post-multiply this vector by

$$\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ \sqrt{3/2} & \sqrt{3/2} \end{bmatrix},$$

yielding the *alr* regressor function

$$9.70 - 2.74 \log(depth), \quad 4.80 - 1.10 \log(depth),$$

in conformity with the direct *alr* approach. Again the [*sand, silt, clay*] residuals are identical to those already determined as in Fig.5.2.

Each of these methods gives exactly the same results. In particular a measure of total variability of the sediments, prior to any regression analysis and as determined by the appropriate measure for each transformation, is 2.47. In each analysis the total variability of the residuals is 0.71, the amount of variability explained by the regression is 1.76, so that the proportion of the variability explained by each of the methods is 1.76/2.47 = 0.71.

I would suggest that the simplest interpretation is provided by *alr* transformations. As depth increases, sand gives way to silt and more so to clay, with differential effects decreasing with depth.

5.2 Fallacy 6a. There are better transformations than logratios for the analysis of compositional data

There has been advocacy of the use of the sphere as a reasonable sample space for compositional data and the established methodology of directional data. See for example Stephens (1982) who converts the unit-sum constraint on composition *x*, namely $x_1 + \cdots + x_D = 1$, by a square route transformation $y_i = \sqrt{x_i} \quad (i - 1, \ldots, D\}$, to a restriction of the *y*-vector to the *D*-dimensional unit sphere. Although Stephens manages to use the spherical von Mises distribution to give a reasonable discrimination between two groups of students, there is a fundamental flaw in the transformation in that the simplex and the sphere are topologically completely unrelated.

Stanley (1990) goes even further away from simplicity, first going to the unit sphere and then continuing by further transformation to spherical polar coordinates. Karl Pearson would have had a field day asking what relationship the angles studied after this transformation had with any reality.

5.3 Confusion between logratio and logarithmic transformations

A simple point is all that is necessary here. As already indicated, Gower (1987) in Section 5 above appears to believe that transformation methodology of compositional data analysis uses a *logarithmic* transformation $z_i = \log x_i \quad (i = 1, \ldots, D)$. This, of course, only changes the unit-sum constraint on the compositional *x*-vector to a more complicated constraint $\exp(z_1) + \cdots + \exp(z_D) = 1$ and achieves no insight into the compositional data.

5.4 Summing up on transformation methodology

To sum up, I see my original transformation approach to compositional data analysis as fundamentally sound. This is not to say that other approaches may give different, possibly better, insights into the description and interpretation of compositional variability.

Another personal comment here. All the early steps in the building of a methodology for compositional data analysis were intuitive, as I think has been the case in many advances in statistics. The formalisation of perturbation, powering, metric, inner product into an undoubtedly elegant geometric structure on S^D should not deceive us into thinking that this is the only structure relevant to resolving problems of compositional data analysis. It certainly plays a substantial role but, as we shall see, it is by no means the only approach required in the discipline.

6. Confusions, fallacies and misunderstandings on the relation between pure mathematics and statistical modelling and analysis

6.1 The relationship of pure mathematics and statistical science (informatics)

Pure mathematics is a deductive activity. From definitions of entities of interest, axioms or rules concerning the behaviour of these entities, and following certain rules of logic, deductions such as lemmata or theorems are made, hopefully mathematically interesting and often elegant consequences are arrived at. An excellent example is the route to the Hilbert space structure of the unit simplex in Section 4.3 above. The simple rules of perturbation, powering, metric and inner product fit together to ensure the Hilbert space property of the unit simplex.

Statistics on the other hand is an inductive activity. Its interest is in dealing with uncertainty and variability. From a sample randomly selected in some way from a population, what can be inferred about the population? How can information consisting of 8-dimensional vectors of clinical measurements on 31 patients, post-operatively found to have two forms, A and B, of an adrenal syndrome, 11 with A, 20 with B, be used to infer the diagnostic form of a new patient whose vector of clinical information has been determined?

The answer to such inductive problems clearly depends on statistical modelling, which essentially uses mathematical ideas in an attempt to describe in probabilistic terms how the variable data may have arisen, and to make inferences from what can be regarded as a reasonable working model. In this activity pure mathematics is surely the servant of the statistical problem. Too often it seems to have become the master and at times an indifferent, even an absurd, master. In the 1950s, when matrix algebra was becoming familiar in some disciplines, and automatic computing was making matrix inversion of modest order a possibility, I was asked by my econometrics boss to invert a matrix of order 6 on EDSAC 1. The matrix consisting of prices which moved in so proportional a way (implying linear dependence of the matrix) that EDSAC hooted at me to indicate ill-conditioning. I tried to explain the folly of attempting to use such a model, but no, the computing staff in the department were set to apply a successive bordering technique to manually invert the matrix to N places of decimals. It was useless to explain that a change in the fourth significant digit in any of the prices would give a completely different inference. The econometrician went on to receive the Nobel Prize for Economics, hopefully not for this piece of work.

I cite this example for the very important reason that there are among us some compositional data analysts who see pure mathematics as the master in statistical modelling of compositional problems. What seems to have happened is that they see the Hilbert space structure of the unit simplex as such an elegant piece of pure mathematics that its use, and only its use, must be the source of solutions to all compositional data analysis. In this approach I think they are totally wrong and I will attempt to explain why in the discussion of a number of fallacies in the next few sections. Part of their approach seems to be to detect differences between results of transformation methodology and staying-in-the-simplex methodology, with, of course, the staying methodology found to be superior. Some of these differences are the results of staying enthusiasts asking the wrong questions about the situation. I emphasize that

transformation and staying methodologies applied properly to a given relevant question produce identical inferences.

The remainder of this section is thus devoted to a series of fallacies involving the Hilbert space simplex as a driving master. Most of these fallacies are the results of ignoring the Jeffreys advice quoted in Section 0.3, substituting an irrelevant question to suit the pure mathematics. A recent excellent television documentary *The Dali Dimension* analysed Dali's considerable understanding and use of mathematical and scientific ideas. The programme involved a confrontation between two commentators. I was certainly on the side of the commentator who said that pure mathematics had no reality; it was only when used in an appropriate modelling situation in relation to some specific question that it touched reality.

6.2 Fallacies and confusions relating to orthonormal coordinates and isometric transformations

The Hilbert space associated with the D-dimensional unit simplex S^D is a $(D-1)$-dimension space and it is easy to construct a set of $D-1$ orthonormal vectors as a basis for the representations of compositions in terms of coordinates with respect to such an orthonormal basis (Aitchison et al., 2001). A number of papers (Egozcue and Pawlowsky-Glahn, 2004; Egozcue et al., 2004; Pawlowsky-Glahn, 2003; Pawlowsky-Glahn and Egozcue, 2001, 2002) have advocated that the perfect way to model a compositional data problem is in terms of such an orthogonal coordinate vector

$$z = [z_1, \ldots, z_{D-1}] = [f_1(x), \ldots, f_{D-1}(x)]. \tag{6.1}$$

The use of these coordinates is equivalent to a transformation which is isometric in the sense that if u and v are the coordinates associated with compositions x and y then the simplicial metric $\Delta_S(x, X)$ is equal to the Euclidean metric in R^{D-1}, namely

$$\sqrt{\sum_{i=1}^{D=1}(u_i - v_i)^2}\ . \tag{6.2}$$

Such transformations have been termed *isometric logratio* abbreviated to *ilr*. All this is admirable and provides simple theoretical results, particularly in establishing least square type properties of certain estimators. Where things go wrong is an implicit belief that this is the only safe way to tackle compositional problems. In my view such an approach to a practical problem is fraught with difficulties. The coordinates in any *ilr* transformation are necessarily a set of orthogonal logcontrasts. So imagine a consultation with a urologist who has analysed the 4-part [1, 2, 3, 4] composition of renal calculi extracted from male patients at operation. The patients subsequently have been classified as of type R (repeater) or type S (single episode patient). The urologist now consults an *ilr* compositional data analyst to ask if there is any way of deciding whether a new patient, whose extracted kidney stone has composition $[x_1, x_2, x_3, x_4]$, is more likely to be R rather than S. So the *ilr* analyst explains to the urologist that it is necessary to examine an ilr transformation, say

$$[(\log x_1 - \log x_2)/\sqrt{2},\ \ (\log x_1 + \log x_2 - 2\log x_3)/\sqrt{6},\ (\log x_1 + \log x_2 + \log x_3 - 3\log x_4)/\sqrt{12}] \tag{6.3}$$

in order to answer his problem. I know a number of urologists and I cannot imagine any of them understanding why such an elaborate transformation would be necessary. And why, they would sensibly ask, is this particular *ilr* transformation relevant. Why not another one? And an even greater stretch of the imagination would be required to envisage them bringing further compositional problems for analysis.

Indeed, ensuring isometry has little to do with this compositional problem. A simple model would be a logistic normal regression of type (R,S) with the regressor function a logcontrast of the 4-part composition, along the lines of the analysis of limestone differences by Thomas and Aitchison (1998), which opens up the possibility of lattice testing to discover if any subcomposition is suitable for the purpose.

An argument from the advocates of the *ilr* coordinates approach is that the *alr* transformation to R^{D-1} is not an isometry. This is true but unimportant since, if there is any requirement to use metric arguments in the transformed space, all that is required is to use the simplicial metric in its *H* form as in (4.9). Indeed, often in parametric modelling the appropriate metric is not the simplicial metric but the 'Mahalanobis metric' of the form

$$(alr(x)-\mu)\Sigma^{-1}(alr(x)-\mu)^T\ . \tag{6.4}$$

Also, in dealing with problems in the simplex where the data vectors are probabilities, the appropriate measure of distance between two vectors *x* and *y* may more appropriately be some form of the Kullback-Leibler (1951) divergence. See Aitchison (1974, 1981, 1985), Aitchison and Kay (1973, 1975), Aitchison, Kay and Lauder (2004), Taylor, Aitchison and McGirr (1971) for some interesting applications of this divergence in assessing the ability of subjects to make inferences.

At CodaWork'03 and CoDaWork'05, when countering this insistence on the use of *ilr* transformations, I said I would look forward to a convincing practical use of the method. As far as I know there has been no progress in demonstrating its applicability.

A major difficulty with the *ilr* transformation approach is that practical compositional situations seldom, if ever, seem to have a natural modelling in terms of a sequence of orthogonal logcontrasts. This is in contrast to the practical use of othogonality in the use of the simplicial singular value decomposition along lines similar to principal component analysis in R^D or principal logcontrast analysis in compositional data analysis (Aitchison, 1983, 1986 Sections 8.3, 8.4). Any $N \times D$ compositional data matrix *X* with *n*th row composition x_n can be decomposed in a power-perturbation form as follows

$$x_n = \xi \oplus (u_{n1}s_1 \otimes \beta_1) \oplus \cdots \oplus (u_{nR}s_R \otimes \beta_R)\,, \tag{6.5}$$

where □ is the centre of the data set, the *s*'s are positive 'singular values' in descending order of magnitude, the □'s are compositions, *R* is a readily defined rank of the compositional data set and the *u*'s are power components specific to each composition. In a way similar to that for data sets in R^d we may consider an approximation of order $r < R$ to the compositional data set given by

$$x_n = \xi \oplus (u_{n1}s_1 \otimes \beta_1) \oplus \ldots \oplus (u_{nR}s_R \otimes \beta_R) \tag{6.6}$$

Such an approximation retains a proportion

$$x_n = \xi \oplus (u_{n1}s_1 \otimes \beta_1) \oplus \ldots \oplus (u_{nR}s_R \otimes \beta_R) \tag{6.7}$$

of the total variability of the $N \times D$ compositional data matrix as measured by the trace of the centred logratio covariance matrix or equivalently in terms of total mutual distances as

$$(N(N-1))^{-1}\sum_{m<n}^{D}\Delta_S^2(x_m, x_n). \tag{6.8}$$

Use of the singular value decomposition is a form of data snooping, getting the data to suggest possible hypotheses. Sometimes it can, in a way similar to such graphical techniques as biplots, be useful in providing convincing evidence of already discovered connections. See, for example, Thomas and Aitchison (1998).

We may also note here that the power-perturbation expression of the singular value decomposition has exactly the same form as regression of a composition on some set of variables. The form is exactly what would be obtained if the logratio form of regression analysis in Aitchison (1986, Chapter 7) were transformed back into terms of the simplex.

6.3 The sub-fallacy of balancing to retain orthogonality

To make my point here about this pure-mathematically mindedness in compositional modelling I reproduce here an extract from my talk at CoDaWork'03. It illustrates how the Hilbert space view of the simplex may not be the appropriate one.

> Given the elegance of the algebraic-geometric (Hilbert space) structure of the simplex it is easy to fall into the pure-mathematical trap that all compositional problems must depend on this structure, that all statistical problems should be addressed in terms of coordinates associated with orthonormal, isometric bases, that orthogonality is closely associated with statistical independence. Let me say here that I think that many of these ideas are important in establishing useful results. For example, such a structure is obviously central to establishing the counterparts of the well known Markov least squares theory associated with R^D. But while we may recognise the simplex as a compositional sample space we must ensure that the ways we place probability measures or distributions on that sample space are appropriate to the applied compositional problem we face. I take an example similar to the statistician's day problem in Aitchison (1986, Sections 1.9, 10.3) for illustrative purposes. Time budgets have become a regular source of information in analysing behaviour patterns in many disciplines. Our example concerns the behaviour pattern of the lesser goilbird, a garden bird whose territory is confined to a particular garden. Its four activities (feeding, fighting | perching, sleeping) divide themselves into two natural divisions: active, including feeding and fighting, and passive, including perching and sleeping. Obvious behavioural questions are whether active and passive patterns are independent and whether these patterns are independent of the division of the day between active and passive.
>
> The time budgets of 60 goilbirds have been observed in 60 gardens over random days.
>
> In terms of the generic composition $[x_1\ x_2\ x_3\ x_4]$ we are here dealing with a partition $[x_1\ x_2 \mid x_3\ x_4]$ of order 1. The relevant question in terms of logratios is whether
>
> $$y_1 = \log(x_1 / x_2),\ \ y_2 = \log(x_3 / x_4),\ \ y_3 = \log\{(x_1 + x_2) / (x_3 + x_4) \quad (6.9)$$
>
> are distributed independently.
>
> Now it has been put to me that amalgamation is not a proper operation in the simplex, presumably because it has no role to play in the Hilbert space structure, and I could have cited this as yet another fallacy:
>
> Fallacy 6c. Amalgamation is not a proper operation within the simplex.
>
> This fallacy leads to tackling such problems by considering an isometric logratio transformation, acknowledging that an appropriate representation of the composition is in terms of coordinates with respect to an orthonormal basis, resulting in
>
> $$x = \log(x_1 / x_2) \otimes e_1 \oplus \log(x_3 / x_4) \otimes e_2 \oplus \log(x_1 x_2 / x_3 x_4) \otimes e_3, \quad (6.10)$$
>
> even suggesting that establishing that
>
> $$z_1 = \log(x_1 / x_2),\ \ z_2 = \log(x_3 / x_4),\ \ z_3 = \log(x_1 x_2 / x_3 x_4)_3 \quad (6.11)$$
>
> are independent would imply independence of y_1, y_2, y_3. This is simply not true, as our data set will demonstrate.

The correlation matrices of y_1, y_2, y_3 and z_1, z_2, z_3 are as follows

1.0000	-0.0022	-0.0861
-0.0022	1.0000	-0.2457
-0.0861	-0.2457	1.0000

and

1.0000	-0.0022	-0.6227
-0.0022	1.0000	-0.6404
-0.6227	-0.6404	1.0000

demonstrating clearly that there is independence associated with the real question, whereas the pseudo-question suggests dependence between the subcompositions and the partition.

Another line of the orthonormalists is that the appropriate modelling must indeed be in terms of the orthonormal coefficients z_1, z_2, z_3 and then it is simply a case of expressing the relevant variables y_1, y_2, y_3 in terms of these coordinates. The first two relations are obviously straight forward but

$$y_3 = \frac{\exp\{\tfrac{1}{2}(z_1 + z_2 + z_3) + \exp\{\tfrac{1}{2}(z_1 - 3z_2 + z_3)}{\exp(z_2) + 1}. \qquad \textbf{(6.12)}$$

This will, of course, lead to a correct analysis but my point is why go to all this complexity, not addressing the problem of interest in its simplest terms. Statisticians have over the past century addressed problems of statistical independence correctly without being aware of any algebraic-geometric structure of their sample spaces. My complaint is not that such structure is unimportant, but that we must not let pure-mathematical ideas drive us into making the statistical modelling more complicated than is necessary. Simplicity in modelling is important, particularly when we have to explain the inferences to less numerate colleagues.

6.4 The fallacy that the coda-dendrogram is a useful exploratory tool

At CoDaWork'05 we met the coda-dendrogram, purporting to be a useful exploratory tool in compositional data analysis. The problem here is still the insistence on predetermined orthogonality transformations, described as sequential binary partitions. Each of these requires a balance which is essentially a logratio of geometric means of components, which, as far as I am aware, has nothing to do with any question which has been posed about the real situation. A typical failure to address the real problem was, if my memory is correct, to address a benchmark problem involving questions of differential diagnosis from blood compositions.

6.5 The fallacy that transformation techniques and staying in the simplex can lead to substantially different inferences.

This fallacy seems to have arisen because of confusions over density functions within spaces with different algebraic geometric structures. It is worth asserting here that density functions are little more than a means to an end, the computation of probabilities of events. Current practice, for example with the logarithmic transformation, in which the transformed variable $y = \log x$ is regarded as distributed as $N(\mu, \sigma^2)$ is to say that the density function of the original variable x is

$$p(x \mid \mu.\sigma) = \frac{1}{(2\pi)^{1/2} x} \exp\left(-\frac{1}{2}\left(\frac{(\log x - \mu)^2}{\sigma^2} \right) \right) \qquad (7.1)\ .$$

Of course this density function is with respect to the Lebesgue measure on the real line. The tradition of including the Jacobian of the transformation is, I believe, to remind transformers, particularly students who often get it wrong, that the Jacobian plays a role in probability computations. For example, the computation here for determining the probability that $a < x < b$ would proceed as

$$P(a < x < b) = \int_a^b p(x \mid \mu, \sigma)dx = \int_{\log a}^{\log b} \phi(y \mid \mu, \sigma)dy = \Phi\left(\frac{\log b - \mu}{\sigma}\right) - \Phi\left(\frac{\log a - \mu}{\sigma}\right) \quad . \tag{7.2}$$

The fuss about density functions seems to be that if one transfers the *x*-density function (including the Jacobian) to the original space it produces isodensity contours which are different from those which would be produced by recognising that the original space has a natural measure different from the Lebesgue measure and so the appropriate density function would take the form

$$q(x \mid \mu.\sigma) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}\left(\frac{(\log x - \mu)^2}{\sigma^2}\right)\right) . \tag{7.3}$$

This is certainly the case, and the same will be true with logratio transformations and logistic-normal distributions in relation to the simplex, but so what. What purpose do such isodensity contours provide? I know of no practical statistical question in compositional data analysis where such distinctions would arise. Here the transformation methodology prevents such confusions. For example, an $N \times D$ compositional data set, with typical composition x transformed by $y = alr(x)$ to R^{D-1} will provide either an estimative normal or predictive Student fit in R^{D-1}. In R^{D-1}, it is known that regions of minimum area (Lebesgue measure), whose normal or Student probability is given, say *P*, are ellipsoids of the form $(y - \hat{\mu})\hat{\Sigma}^{-1}(y - \hat{\mu})^T = c$, and *c* is easily determined to accommodate the preassigned *P*. Such an approach is reliable and can be found, for example, in Aitchison (1986, Section 7.10).

My fear here is that this pure mathematical fussiness about what is and is not a density function leads only to confusion and seeks to find inference differences between transformation methodology and staying-in-the-simplex methodology where in reality none exists. I repeat my earlier comment that any density function is a tool to compute probabilities of events of interest. As long as probabilities are computed properly inferences will agree.

7. Conclusions and discussion

There are many important aspects of compositional data analysis which have been omitted from this paper, for example the persistent and demanding problem of modelling situations with essential zero components, the question of whether the skew logistic-normal class of distributions on the simplex (Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999; Mateu-Figueras et al. 1998) has parameterisation suited to simple investigation of compositional hypotheses, questions of the development of regionalised (spatial) compositions along the lines of Pawlowsky (1986), Pawlowsky et al. (1995), the underuse of existing tools (Aitchison and Thomas, 1998; Aitchison and Barceló-Vidal, 2002) for the exploration of compositional processes, including compositional time series, the wide range of problems involving convex linear combinations of compositions.

This paper may appear adversarial in many ways. This has been a deliberate device in the hope that it may lead to vigorous and, as usual at CoDa worshops, friendly discussion on how the statistical modelling tools now at our disposal may be best applied to answer real questions from the many disciplines where compositional data arise; and where these tools fail, how we may devise alternatives which are appropriate to dealing with the new situation.

A main conclusion is that there still remain many fallacies, confusions and misunderstandings in compositional data analyses. Of these the ones that concern me most are

(a) the continuing use of standard R^D multivariate statistical analysis;

(b) the obvious continuing misuse of scattergrams of raw components, particularly in student instruction in a number of disciplines;

(c) the increasing use of data-mining, data-snooping, followed by confirmation of suggested hypotheses by analysis of the same data;

(d) the view that every compositional data analysis must fit into the Hilbert space structure of the simplex;

(e) the view that the *alr* transformation is suspect because it is not a direct isometry;

(f) the view that transformation and staying-in-the-simplex methods applied to the same real question can lead to different inferences.

Acknowledgments

My continued interest in compositional data analysis has been largely a contest between what has come to be affectionately known as the Girona Gang, who by all sorts of skilful devices, have kept me working, and my family, who describe me as long past my sell-by date and with whom I largely agree.

References

Aitchison, J. (1974). Hippocratic Inference. *IMA Bulletin* **10**, 48-53.

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

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *J. R. Statist. Soc*.**B44**, 139-177.

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

Aitchison, J. (1981). Some distribution theory related to the analysis of subjective performance in inferential tasks, in *Statistical Distributions in Scientific Work* (Taillie, C., Patil, G.P. and Baldessari, B., eds), Vol 5, pp.363-385. Dordrecht, Holland: D. Reidel Publishing Company.

Aitchison, J. (1985). Practical Bayesian problems in simplex sample spaces. In *Proceedings of the Second International Meeting on Bayesian Statistics* (Barnardo, J.M., DeGroot, M.H., Lindley, D.V. and Smith, A.F.M., eds). pp.15-31.

Aitchison, J. (1986a). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall. Reprinted in 2003 by The Blackburn Press.

Aitchison, J. (1986b). CODA: *A Software Package for Compositional Data* (with diskette). London: Chapman and Hall.

Aitchison, J., (1989). Letter to the Editor. Measures of location of compositional data sets. *Math. Geology* **21**, 787-790.

Aitchison, J. (1990a). Letter to the Editor. Comment on "Measures of Variability for Geological Data" by D. F. Watson and G. M. Philip. *Math. Geology*. **22**, 223-226.

Aitchison, J. (1990b). Relative variation diagrams for describing patterns of variability of compositional data. *Math. Geology* **22**, 487-512.

Aitchison, J. (1991a). Letter to the Editor. Delusions of uniqueness and ineluctability: *Math. Geology*

23, 275-277.

Aitchison, J. (1991b). A plea for precision in Mathematical Geology. *Math, Geology* **23**, 1081-1084.

Aitchison, J. (1992). On criteria for measures of compositional differences. *Math Geology* **24**, 365-380.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is easy. In *Proceedings of the Third Annual Conference of the International Association for Mathematical Geology* (ed. Vera Pawlowsky Glahn). 3-35. Barcelona: CIMNE

Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Math. Geology* **31**, 563-589.

Aitchison, J. (2001). Simplicial inference. In *Algebraic Methods in Statistics and Probability* (M. A. G. Viana and D. St. P. Richards, eds), pp. 1-22. Contemporary Mathematics Series 287. Providence, Rhode Island: American Mathematical Society.

Aitchison, J. and Bacon-Shone, J.H. (1984). Logcontrast models for experiments with mixtures. *Biometrika* **71**, 323-330.

Aitchison, J. and Barceló-Vidal, C. (2002). Compositional processes: a statistical search for understanding. In *Proceedings of the Eighth Annual Conference of the International Association for Mathematical Geology* 3, pp. 381-386.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. Math. Geology 32, 271-275.

Aitchison, J., Barceló-Vidal, C., Egozcue, J. J. and Pawlowsky-Glahn, V. (2002). A concise guide to the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In Terra Nostra : *Proceedings of the Eighth Annual Conference of the International Association for Mathematical Geology*, pp. 387-392..

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2001). Reply to Letter to the Editor by S. Rehder and U. Zier on 'Logratio analysis and compositional distance' by J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández and V. Pawlowsky-Glahn. *Math. Geology* **33**, 849-860.

Aitchison, J., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (2002). Some comments on compositional data analysis in archaeometry, in particular the fallacies in Tangri and Wright's dismissal of logratio ,analysis. *Archaeometry*, **44**, 295-304...

Aitchison, J. and Greenacre, M. (2002). Biplots for compositional data: *Appl. Statist.* **1**, 375-382.

Aitchison, J. and Kay, J.W. (1973). A diagnostic competition. *IMA Bulletin* **9**, 382-383.

Aitchison, J. and Kay, J.W. (1975). Principles, practice and performance in decision-making in clinical medicine. In *Proceedings of the 1973 NATO Conference on The Role and Effectiveness of Decision Theories in Practice,* eds K.C. Bowen and D.J. White. London: English Universities Press.

Aitchison, J., Kay, J.W. and Lauder, I.J. (2004). *Statistical Concepts and Applications in Clinical Medicine.* London: Chapman and Hall/CRC.

Aitchison, J. and Thomas, C. W. (1998) Differential perturbation processes: a tool for the study of compositional processes. In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology* (A. Buccianti, G. Nardi and R. Potenza, eds), pp. 499-504. Naples: De Frede.

Aitchison, J., and Shen, S. M. (1980). Logistic-normal distributions: some properties and uses: *Biometrika* **67**, 261-272.

Aitchison, J. and Thomas, C. W. (1998) Differential perturbation processes: a tool for the study of compositional processes. In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology* (A. Buccianti, G. Nardi and R. Potenza, eds), pp. 499-504. Naples: De Frede.

Azzallini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715-726.

Azzallini, A. and Capitanio, A. (1999). Statistical application of the multivariate skew-normal distribution. *J. R. Statist. Soc.* B, **61**, 579-602.

Baxter, M.J., Beardah, C.C., Cool, H.E.M.and Jackson, C.M. (1903). Compositional data analysis in archaeometry. CoDaWork03,

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *J. Amer, Statist. Ass*. **96,** 1205-1214.

Chayes, F. (1960). On correlation between variables of constant sum. *J. Geophys. Res*., **65,** 4185-4193**.**

Chayes, F. (1962), Numerical correlation and petrographic variation. J. Geology, **70**, 440-452.

Chayes, F. (1971). *Ratio Correlation*. University of Chicago Press.

Chayes, F. and Kruskal, W. (1966). An approximate statistical test for correlation between proportions. *J. Geology*. **74**, 592-702.

Fenner, C. N. (1913). The stability relations of the silica minerals. *Amer. J, Science* **36**, 334-384.

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

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

Gower, J. C.,(1987). Introduction to ordination techniques, in Legendre, P. and Legendre, L., eds., *Developments in Numerical Ecology*: Springer-Verlag, Berlin, pp. 3-64.

Harker, A. (1909). *The Natural History of Igneous Rocks*. New York: Macmillan.

Jeffreys, H. (1961), *The Theory of Probability* (3rd edition). Oxford University Press.

Kullback. S. and R. A. Leibler. (1951), On information and sufficiency. *Ann. Math. Stat.* **22**. 525-540.

Mateu-Figueras. G., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998). Modeling compositional data with multivariate skew-normal distributions. In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology* (A. Buccianti, G. Nardi and R. Potenza, eds), pp. 532-537. Naples: De Frede.

Pawlowsky, V. (1986). *Räumliche Strukturanalyse und Schätzung ortsabhängiger Kompositionen mit Anwendungsbeispeilen aus der Geologie*: unpublished dissertation, FB Geowissenschaften, Freie Universität Berlin.

Pawlowsky, V., Olea, R. A. and Davis, J. C. (1995). Estimation of regionalized compositions: a comparison of three methods. *Math. Geology* **27**, 105-148.

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

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

Pearson, K. (1906). Skew frequency curves. A rejoinder to Professor Kapteyn: *Biometrika*. **5**, 168-171.

Rehder, U. and Zier, S. (2001). Comment on "Logratio analysis and compositional distance by Aitchison et al. (2000)". *Math. Geology* 33, 845-848.

Stanley, C. R., 1990, Descriptive statistics for N-dimensional closed arrays: a spherical coordinate approach: *Math. Geology*. **22**, 933-956.

Stanley, C. R., 1990, Descriptive statistics for N-dimensional closed arrays: a spherical coordinate approach: *Math. Geology* **22**, 933-956.

Stephens, M.A., 1982, Use of the von Mises distribution to analyse continuous proportions: *Biometrika* **69**, 197-203.

Taylor, T.R., Aitchison, J. and McGirr, E.M. (1971). Doctors as decision-makers: computer-assisted study of diagnosis as a cognitive skill. *Brit. Med. J.* **3**, 35-40.

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. In A. Buccianti, G. Nardi and R. Potenza, Eds., *Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology*, De Frede, Naples, pp. 549-554.

Watson, D. F. (1990). Reply to Comment on "Measures of Variability for Geological Data" by D. F. Watson and G. M. Philip: *Math. Geology* **22**, 227-231.

Watson, D. F. (1991). Reply to "Delusions of Uniqueness and Ineluctability" by J. Aitchison: *Math. Geology*, **23**, 279.

Watson, D.E. and Philip, G.M. (1989). Measures of variability for geological data. *Math. Geology* **27**, 233-254.

Whitten, E. H. T., 1995, Open and closed compositional data in petrology: *Math. Geology*. **27**, 789-806.

Woronow, A. (1997a). The elusive benefits of logratios. In *Proceedings of the Third Annual Conference of the International Association for Mathematical Geology* (V. Pawlowsky-Glahn, ed.), pp. 97-101. Barcelona: CIMNE.

Woronow, A. (1997b). Regression and discrimination analysis using raw compositional data - is it really a problem?. In *Proceedings of the Third Annual Conference of the International Association for Mathematical Geology* (V. Pawlowsky-Glahn, ed.), pp. 157-162. Barcelona: CIMNE.

Zier, U. and Rehder, S. (1998), Grain-size analysis - a closed data problem. In *Proceedings of the Fourth Annual Conference of the International Association for Mathematical* Geology (A. Buccianti, G. Nardi and R. Potenza, eds), pp. 555-558. Naples: De Frede.

Table 1. Three part [1, 2, 3] compositions

1	2	3
0.16	0.48	0.36
0.43	0.27	0.30
0.16	0.74	0.10
0.69	0.15	0.16
0.91	0.02	0.07
0.80	0.13	0.07
0.58	0.03	0.40
0.46	0.30	0.24
0.70	0.14	0.16
0.71	0.15	0.14
0.53	0.34	0.14
0.45	0.38	0.17
0.35	0.34	0.31
0.38	0.26	0.36
0.24	0.66	0.10
0.28	0.48	0.24
0.32	0.26	0.42
0.49	0.35	0.16
0.13	0.58	0.29
0.54	0.26	0.20
0.70	0.20	0.10
0.55	0.23	0.22
0.31	0.41	0.28
0.46	0.19	0.35
0.47	0.33	0.20
0.42	0.34	0.24
0.39	0.13	0.48
0.83	0.03	0.14
0.70	0.21	0.09
0.51	0.23	0.26
0.72	0.15	0.13
0.54	0.19	0.27
0.86	0.07	0.07
0.30	0.32	0.38
0.14	0.84	0.02
0.68	0.22	0.10
0.82	0.04	0.14
0.35	0.52	0.13
0.74	0.14	0.12
0.87	0.08	0.05

Table 2. Time budgets of 50 goilbirds

feeding	fighting	perching	sleeping
0.5476	0.0107	0.0113	0.4303
0.5385	0.0253	0.0090	0.4271
0.4712	0.0175	0.0211	0.4902
0.4830	0.0091	0.0553	0.4526
0.4340	0.0031	0.1003	0.4627
0.5220	0.0169	0.0321	0.4290
0.5939	0.0027	0.0115	0.3919
0.5781	0.0229	0.0222	0.3767
0.4733	0.0047	0.0122	0.5098
0.4863	0.0309	0.0096	0.4732
0.5277	0.0220	0.0058	0.4445
0.4440	0.0128	0.0044	0.5389
0.5106	0.0076	0.0215	0.4603
0.5264	0.0016	0.0406	0.4313
0.5323	0.0088	0.0262	0.4327
0.4396	0.0119	0.0258	0.5227
0.5981	0.0067	0.0191	0.3761
0.5453	0.0312	0.0121	0.4115
0.3141	0.0063	0.1560	0.5236
0.4096	0.0049	0.0227	0.5628
0.4630	0.0112	0.0068	0.5190
0.3388	0.0073	0.0235	0.6304
0.6120	0.0095	0.0107	0.3679
0.5121	0.0063	0.0205	0.4611
0.5489	0.0020	0.0149	0.4341
0.4105	0.0011	0.0129	0.5755
0.5107	0.0048	0.0046	0.4798
0.5914	0.0396	0.0116	0.3574
0.5500	0.0071	0.0050	0.4378
0.5452	0.0171	0.0190	0.4186
0.5218	0.0257	0.0477	0.4048
0.4907	0.0046	0.1617	0.3429
0.4085	0.0047	0.0442	0.5425
0.6490	0.0143	0.0231	0.3136
0.3846	0.0101	0.0721	0.5333
0.5142	0.0218	0.0323	0.4317
0.4805	0.0504	0.0682	0.4009
0.6062	0.0520	0.0137	0.3281
0.4494	0.0251	0.0280	0.4975
0.5978	0.0162	0.0100	0.3759
0.4533	0.0070	0.0128	0.5269
0.5091	0.0075	0.0133	0.4701
0.5280	0.0314	0.0428	0.3978
0.4216	0.0040	0.0290	0.5454
0.5417	0.0066	0.0039	0.4478
0.6328	0.0029	0.0801	0.2842
0.4924	0.0146	0.0418	0.4512
0.6818	0.0126	0.0035	0.3021
0.4337	0.0131	0.0186	0.5346
0.7006	0.0065	0.0167	0.2762
0.4954	0.0032	0.0118	0.4895
0.5156	0.0059	0.0206	0.4579
0.4277	0.0006	0.0367	0.5350
0.3431	0.0073	0.0761	0.5734
0.4692	0.0057	0.0068	0.5183
0.4886	0.0578	0.0083	0.4453
0.5483	0.0169	0.0114	0.4234
0.3339	0.0367	0.0348	0.5946
0.3455	0.0070	0.0980	0.5495
0.4376	0.0279	0.1273	0.4072