*Mathematical Geology, Vol. 31, No. 5, 1999*

# Logratios and Natural Laws in Compositional Data Analysis[1]

**John Aitchison**[2]

*The impossibility of interpreting correlations of raw compositional components and associated statistical methods has been clearly demonstrated over the last four decades and alternative statistical methodology developed. Despite this a return to the ''traditional'' use of raw components has been advocated recently and alternative methodology such as logratio analysis strongly criticized. This paper exposes the fallacies in this recent advocacy and demonstrates the constructive role that logratio analysis can play in geological compositional problems, in particular in the investigation of natural laws and in subcompositional investigations.*

**KEY WORDS:** logcontrast principal component analysis, raw correlation misinterpretation, subcompositional hypotheses, subcompositional coherence, variability of olivines.

## INTRODUCTION

This paper is concerned with important statistical issues involved in the analysis and interpretation of compositional datasets, such as major oxide and trace element compositions of rocks and sedimentary compositions. Such datasets inevitably display variability and so require a statistical methodology appropriate to the special nature of compositions—the so-called constant-sum property—to allow meaningful interpretation of the nature of this variability and the consequent geological inferences. In two recent papers presented at IAMG97, Woronow (1997a,b) explicitly and implicitly rejects the warnings of Pearson (1897), Chayes (1949, 1960, 1962, 1971), Sarmonov and Vistelius (1959), Krumbein (1962), Mosimann (1962, 1963), Chayes and Kruskal (1966), Aitchison (1981, 1982, 1986, 1997), Le Maitre (1982), Davis (1986), Pawlowsky (1986), Rock (1988), Woronow (*sic*, 1987),

[1] Received 28 April 1998; accepted 17 August 1998.

[2] Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK. e-mail: John.Aitchison@btinternet.com

0882-8121/99/0700–0563$16.00/1 © 1999 International Association for Mathematical Geology

Woronow and Love (1990), Reyment and Savazzi (1999), and many others that in compositional data analysis there is no meaningful way to interpret correlations of raw components and associated forms of multivariate statistical analysis designed for unconstrained data. Not only so, but he exhorts geologists to ignore most of the more promising developments of the last two decades for effective compositional data analysis—in particular what has come to be known as logratio analysis. These views were vigorously challenged at IAMG97, but no record of the ensuing criticisms of Woronow's dismissal of logratio analysis exists. This is an account to put the record straight by exposing the many fallacies and misstatements in the Woronow (1997a,b) papers, and by so doing to reemphasize the advantages of designing appropriate statistical analysis suited to the nature of the objects studied. We confine attention to questions truly compositional in nature and not to the irrelevancies of the well-known distinction between unconstrained data (vectors in $R^D$) to which standard multivariate analysis is appropriate and compositional data, which require a completely different methodology.

Woronow's disparagement of logratio analysis of compositions separates into a number of common misunderstandings of the nature of logratio analysis, and we shall take these individually in a logical sequence against the background of Woronow's statements and his illustrative examples.

## THE NATURE OF LOGRATIO ANALYSIS

Woronow (1997a, p. 99) makes the following general statement:

(a) ‘‘Logratioing accomplishes one aim. It creates a new set of variables that can exhibit mutual independence.’’

This is only a quarter truth. The purpose of logratioing is to supply a meaningful, interpretable description of the interdependence of components of compositions free from all the fallacious interpretations that emerge from raw component analysis. As we shall see later, these are traps that Woronow falls into in his illustrative examples. The rationale and relevance of the logratio covariance and correlation structures have been presented at length in many publications, for example and most recently in Aitchison (1997), and will not be reargued here. Suffice it to recall that the logratio covariance structure has the essential property for compositional data analysis of subcompositional coherence: logratio covariances and correlations within a subcomposition are identical to those within the full composition. In the traditional jargon of open and closed sets, the logratio covariance structures are identical in the open and closed sets.

That this property does not hold for raw component correlations immediately rules out raw component analysis as a viable tool for intelligent discussion of compositional variability.

Statement (a) of Woronow (1997a) is followed, again on p. 99, by:

(b) "The transformation cannot add information, therefore it cannot expand the scope of questions that can be resolved with compositional data, . . ."

with a reinforcement of this on p. 101:

(c) "Although logratioing creates variables with the potential for mutual independence, this or other transformations do not expand the breadth of questions that compositional data can address, and transformation may actually reduce that breadth."

Statement (b) is, of course, true, but does not note that equally the logratio transformation *does not lose* any information. This is so because there is a one-to-one correspondence between any $D$-part composition $(x_1, . . . , x_D)$ and its logratio vector $(y_1, . . . , y_{D-1})$, the two transformations being

$$y_i = \log(x_i/x_D) \ (i = 1, . . . , D - 1) \tag{1}$$

$$\begin{aligned} x_i &= \exp(y_i)/\{\exp(y_1) + \cdot \cdot \cdot + \exp(y_{D-1}) + 1\} \ (i = 1, . . . , D - 1) \\ x_D &= 1/\{(\exp(y_1) + \cdot \cdot \cdot + \exp(y_{D-1}) + 1\} \end{aligned} \tag{2}$$

This means that any statement about the raw components of a composition can be expressed as an equivalent statement in terms of logratios and equally any statement in terms of logratios can be expressed as an equivalent statement in terms of raw components. To claim therefore as in (c) that the logratio transformation may reduce the breadth of problems that compositional data can address is obviously absurd.

The essential feature of statistical investigation of compositional data in terms of logratios is thus that, *without any loss of information* about compositional variability, the way is open to study *any* statement or hypothesis about the nature of compositional variability free from the known fallacies of raw compositional data analysis.

## LOGRATIO LAWS IN NATURE

Following statement (b), Woronow on p. 99 continues:

(d) "Given this fact, one must ask what logratioing actually contributes to the testing of or discovery of natural lawa and causal relationships," and later on p. 100 with:

(e) “Nowhere in nature is a logratio-mixing law known.”

These seem remarkable statements about natural laws, particularly by a geologist. Even a statistician attempting to learn more about geology meets in an elementary book on geochemistry (Krauskopf, 1979) as early as page 5 a logratio law, in a first example to illustrate the nature of an equilibrium constant $K$. There we learn that in the reaction

$$H_2 + Cl_2 \rightleftarrows 2HCl$$

the corresponding equilibrium constant $K$ is given by

$$K = \frac{[HCl]^2}{[H_2][Cl_2]}$$

where the components in brackets are expressed in appropriate units of concentration. This is nothing more nor less than an equivalent logratio condition,

$$\log([H_2]/[HCl]) + \log([Cl_2]/[HCl]) = -\log(K)$$

or equivalently and more symmetrically as a logcontrast condition,

$$\log([H_2]) + \log([Cl_2]) - 2\log([HCl]) = -\log(K) \tag{3}$$

We note here that a logcontrast of a composition $(x_1, \ldots, x_D)$ is of the form

$$\beta_1 \log x_1 + \cdots + \beta_D \log x_D \text{ with } \beta_1 + \cdots + \beta_D = 0 \tag{4}$$

which can always be expressed in terms of logratios, for example as

$$\beta_1 \log (x_1/x_D) + \cdots + \beta_{D-1} \log (x_{D-1}/x_D)$$

Indeed, in another IAMG97 paper Woronow (1997c) himself uses such concepts, expressible as simple logratio relationships, in studying equilibrium liquid lines.

For the general form of reaction

$$a[A] + b[B] + \cdots \rightleftarrows y[Y] + z[Z] + \cdots$$

the corresponding equilibrium constant

$$K = \frac{[Y]^y[Z]^z \ldots}{[A]^a[B]^b \ldots}$$

has an equivalent logarithmic form:

$$\log K = y\log([Y]) + z\log([Z]) + \cdots - a\log([A]) - b\log(\{B]) - \cdots$$

Although the coefficients of such relationships in general do not satisfy the logcontrast condition $y + z + \cdots - a - b - \cdots = 0$ and while concentrations [in brackets] are not compositions, the logarithmic version encourages the view that one sensible way to identify patterns in *compositional* datasets is to search for constant logcontrasts of the components of the compositions. Any such relationship may, of course, be translated back into terms of raw components of the composition and this is likely to be the preferred form for the geologist. The role of logratio analysis is to provide an appropriate tool for identifying such patterns by sound statistical procedures.

Woronow (1997a, p. 101) continues with the following statement:

(f) "Therefore, whenever possible, compositional data should be analyzed within their own framework, as has been done traditionally in geology, chemistry, physics and a variety of other hard and soft sciences. . . ."

There is in this statement an implication that geologists should follow the traditional raw component techniques of other branches of hard and soft sciences and, *a fortiori*, that logratio laws have no relevance in such sciences. Let us record here simply that a deeper knowledge of these sciences may lead to a retraction of this view. A first and obvious example is the fundamental 1908 Hardy–Weinberg Law in genetics, which applied for example to (MM, NN, MN) blood group compositions can be expressed as

$$\mathrm{MM.NN} = 4\ \mathrm{MN}^2$$

or equivalently as the logratio or logcontrast law

$$\log \mathrm{MM} + \log \mathrm{NN} - 2 \log \mathrm{MN} = \log 4$$

An elementary account of this natural logratio law and its derivation from probabilistic axioms of genetics can be found in Edwards (1977, p. 22–24). An illustration of how the law can be inferred from logratio analysis of an actual {MM, NN, MN) compositional dataset is provided in Aitchison (1999). Since the compositional form and the logcontrast form are equivalent, there is no advantage or disadvantage in either. Direct derivation of the compositional form depends on the theoretical development of some probabilistic axioms of genetics (much as stoichiometric principles in geochemistry operate). The logcontrast form arises from a simple application of logratio analysis to an actual compositional dataset in a manner similar to an application to olivines in the next section of this paper. No amount of statistical analysis of *raw* compositional data, which implies consideration of *linear* forms in the raw components, will lead to the Hardy–Weinberg *curve*.

The point here is surely that if the Hardy–Weinberg law had not been deducible from genetic axioms, logratio analysis of actual compositional datasets would have led to the logcontrast form and the translated-back version would almost certainly have led geneticists to formulate the then obvious genetic axioms. Since the processes producing compositional datasets in geology are so often not fully understood such logratio analysis would seem a sensible starting point in any attempt to identify relationships among the components of the composition, from which possible theories of genesis might emerge.

As a second example, in what Woronow may regard as one of the softer sciences, economics, the useful concept of income elasticity of demand in household budget analyses is simply expressible in logratio terms. The problem here involves compositions in the form of household budget patterns, consisting of the proportions of total expenditure devoted to the various commodity groups. An important initial question here is whether the pattern (composition) is independent of income, or total expenditure (size)—equivalent to asking whether all income elasticities of demand are equal. Usually this hypothesis would be rejected, in which case the further logratio analysis leads to the estimation of the elasticities through a logratio form of statistical analysis. Aitchison (1986, Section 9.6) provides a simple example. Such analyses date back to Houthakker (1960).

Since the list of successful applications of logratio analysis could be extended easily to agriculture, industrial science, literary analysis, materials science, medicine, physiology, psephology, psychology, and sociology, we may be tempted to ask why Woronow has not managed to detect successful examples in geology. The answer is perhaps to be found in the quality of the argument he presents in his persistence in the use of an irrelevant and meaningless form of statistical analysis in the following examples in which he attempts to denigrate the concept of logratio analysis.

## LOGRATIO LAWS AND OLIVINES

Let us examine the Woronow (1997a, p. 99) advocacy of raw component analysis in the study of the ternary system of the three-part composition (Fe, Mg, Si) associated with his "ideal olivine":

(g) "Correlations in RCD (raw component data) may directly image underlying order. For instance, in olivine the perfect negative correlation between Fe and Mg both uncorrelated with Si, faithfully recites the mineral's crystal chemistry. The correlations are not anomalies introduced by the constant-sum constraint. Whether an analyst unfamiliar with the concept of stoichiometry would ascribe

> the correlation structure to a solid solution may be questioned. A failure to do so would speak to inherent ambiguities in inductive reasoning, not to problems in analyzing correlations in the compositional data. Therefore, correlations in RCD are not inherently fallacious or ambiguous. A prepared mind would be capable of interpreting the physical causes recorded by such data."

This is in fact the perfect example for illustrating the folly of the raw component argument. The argument sets

$$\mathrm{corr}(\mathrm{Fe}, \mathrm{Mg}) = -1, \mathrm{corr}(\mathrm{Fe}, \mathrm{Si}) = 0, \mathrm{corr}(\mathrm{Mg}, \mathrm{Si}) = 0 \tag{5}$$

The *logical* consequences of Eq. (5), easily deduced from the well-known zero row- and column-sum property of a raw covariance matrix, is that the raw covariance matrix for such three-part olivine compositions must take the form

| | Fe | Mg | Si |
|---|---|---|---|
| Fe | c | −c | 0 |
| Mg | −c | c | 0 |
| Si | 0 | 0 | 0 |

From var(Si) = 0 we see that Si must be constant, so that it follows, whatever the natural law that determines the proportions of Fe and Mg, their sum Fe + Mg will be constant and so corr(Fe, Mg) will be −1. Thus the perfect negative correlation is nothing more than an arithmetic artefact, a consequence of the structure of any such raw correlation matrix and indeed attributable to the constant-sum effect of the (Fe, Mg) subcomposition, and therefore provides absolutely no information about the relationship between Fe and Mg. Such arguments have, of course, been countered many times in the literature cited. Indeed, it is trivial to construct *open* datasets here with, for example, zero or even positive correlations between Fe and Mg, which yield *closed* datasets exhibiting the covariance structure (5). For example, the open data set

| Fe | Mg | Si |
|---|---|---|
| 3.30 | 26.88 | 20.12 |
| 1.38 | 11.42 | 8.53 |
| 4.90 | 34.41 | 26.20 |
| 5.49 | 47.77 | 35.51 |

| Fe | Mg | Si |
|---|---|---|
| 1.99 | 13.41 | 10.27 |
| 10.40 | 123.75 | 89.44 |
| 10.39 | 70.63 | 54.02 |
| 4.36 | 33.19 | 25.04 |

has corr(Fe, Mg) = 0.919, corr(Fe, Si) = 0.932, corr(Mg, Si) = 0.999, whereas the composition formed as the closed set of the above open data has the ideal olivine correlations, corr(Fe, Mg) $= -1$, corr(Fe, Si) $= 0$, corr(Mg, Si) $= 0$.

Rather than continue discussion of this so-called ideal olivine, it seems more constructive to see how logratio analysis can deal with actual olivine compositional datasets in the search for natural laws. We have examined eleven such datasets, as set out in Table 1. A simple logratio technique here is to perform a logcontrast principal component analysis (Aitchison, 1983, 1986, Sections 8.3–4). Applied to the first dataset this produces eigenvalues $\lambda_1$ and $\lambda_2$ and corresponding logcontrasts:

$$\lambda_1 = 0.2622 \qquad -0.809 \log \mathrm{Fe} + 0.501 \log \mathrm{Mg} + 0.308 \log \mathrm{Si}$$

$$\lambda_2 = 0.0093 \qquad 0.111 \log \mathrm{Fe} + 0.645 \log \mathrm{Mg} - 0.756 \log \mathrm{Si}$$

The near zero eigenvalue associated with the second logcontrast implies that this logcontrast is almost constant. Scaling this so that the coefficient of log Si is 1, to allow comparison with other datasets, we have the relationship

$$0.147 \log \mathrm{Fe} + 0.853 \log \mathrm{Mg} - \log \mathrm{Si} = \text{constant}$$

where the constant is estimated from the sample compositions. This can be expressed in a more familiar way, in the form analogous to equilibrium constant forms, as

$$\left(\frac{\mathrm{Fe}}{\mathrm{Si}}\right)^a \left(\frac{\mathrm{Mg}}{\mathrm{Si}}\right)^b = c \tag{6}$$

where $a = 0.147, b = 0.853, c = 0.958$. Table 1 gives the $(a, b, c)$ combinations

**Table 1.** Sources of Olivine Compositional Datasets and Estimated Combinations $(a, b, c)$ in the Relationship $(\mathrm{Fe/Si})^a$ $(\mathrm{Mg/Si})^b$ $= c$

| Source | *a* | *b* | *c* |
|---|---|---|---|
| Eissen and others (1989, Table 2b) | 0.147 | 0.853 | 0.958 |
| Chai and Naldrett (1992, Table 2) | 0.167 | 0.833 | 0.896 |
| Allan and others (1989, Table 3) | 0.369 | 0.631 | 0.720 |
| Beard and Day (1988, Table 2) | 0.291 | 0.709 | 0.842 |
| Fan and Hooper (1989, Table 3) | 0.118 | 0.882 | 0.998 |
| Fan and Hooper (1991, Table 5) | 0.251 | 0.749 | 0.877 |
| Kamenetsky and others (1995, Table 1) | 0.163 | 0.837 | 0.924 |
| Deer, Howie, and Zussman (1982, Table 4) | 0.111 | 0.889 | 1.034 |
| Deer, Howie, and Zussman (1982, Table 5) | 0.089 | 0.911 | 1.013 |
| Deer, Howie, and Zussman (1982, Table 7) | 0.709 | 0.291 | 0.845 |
| Deer, Howie, and Zussman (1982, Table 8) | 0.870 | 0.130 | 0.907 |

for the eleven olivine data sets. Since *a* and *b* are always both positive, we see in these logratio laws the typical Fe–Mg exchange feature of olivines. Increases in Mg are at the expense of Fe and vice versa. The laws are quantitative and similar in structure to those abounding in olivine literature as in Deer, Howie, and Zussman (1982). Is it not reasonable in an experimental or observational science dependent on the analysis of compositional data to explore for laws similar to Eq. (6) above? It certainly seems to an outside observer that for new compositional datasets Eq. (6) is a useful starting point. For example, what stoichiometric considerations are necessary to explain the variation in the $(a, b, c)$ configurations in Table 1? The configurations are significantly different and demand some sort of geological explanation.

Woronow (1997a) cites another ideal chemical reaction involving 4-part compositons (albite, kaliophilite, orthoclase, nepheline) in support of his antilogratio thesis, with corr(albite, kaliophilite) = 0, corr(orthoclase, nepheline) = 0, and corr(albite + kaliophilite, orthoclase + nepheline) = −1 supposedly being the crucial correlations. The argument here is equally fallacious with the perfect negative correlation arising as a logical consequence of the fact that in such a four-part composition albite + kaliophilite = 1 − (orthoclase + nepheline), whatever the natural laws obtaining in the determination of the compositions.

## LOGRATIOS IN HYPOTHESIS TESTING

### The Missing-One-Out Fallacy Revisited

In his second IAMG97 paper Woronow (1997b) attempts to demonstrate that the naive device of omitting one of the components of a compositional vector is a satisfactory approach to compositional problems involving regression and discriminant analysis. He appears to regard the problem as being solely due to the singularity of the raw covariance matrix and on page 158 makes the following statement:

(h) ‘‘The trick that overcomes the singular-matrix problem is trivial—it does not require logratios or any other data transformation—simply delete one compositional variable then execute the analysis.’’

Woronow’s claim to success is that in a regression analysis such as his example of regressing the Easting on Darss Sill granulometric compositional data the same regression results occur whichever of the components is dropped. Indeed, he would have obtained exactly the same results if he had retained all the components and used a pseudo-inverse, such as the

Moore–Penrose inverse, in his analysis. The equivalence is a mathematical tautology, a logical consequence of the singularity of the raw convariance matrix arising from the multicollinearity of the data, not an overcoming of the compositional problem. A simple analogy, devoid of any compositional or constant-sum argument, may help in pinpointing the nature of this equivalence. Suppose that in a woodland survey a sample of trees is measured for height $H$, diameter $D$, and circumference $C$ at a specified height and, after felling, usable volume $V$ of timber. The objective is to try to predict usuable volume given the other three measurements. Let us suppose that the analyst proposes the regression model:

$$V = \alpha + \beta H + \gamma D + \delta C + \text{error}$$

The covariance matrix here of the covariates $H$,$D$,$C$ is singular because of the relationship $C = \pi D$. Despite this, we shall obtain the same regression results whether we drop $D$ or drop $C$ or retain $D$ and $C$, and use a Moore–Penrose inverse in our regression calculations.

What is being forgotten in all this manipulation of singular matrices is that uniqueness of result is not the real criterion of successful regression but the quality of the regression—for example, the reliability of usuable volume prediction based on the covariates. In this example, we are clearly likely to do better by using a different form of regression predictor, taking into account the physical nature of the problem and using a multiplicative model

$$V = \alpha H^{\beta} D^{\gamma} \times \text{error}$$

dropping the superfluous $C$, or equivalently,

$$\log V = \varepsilon + \beta \log H + \gamma \log D + \text{error}$$

where $\varepsilon = \log\alpha$. May it not be the case that improved reliability will be provided if we take account of the special nature of compositions? We can regress not on a reduced set of raw components but on a logcontrast of the components, with a model expressing the response $z$, say, in terms of the $D$-part compositions $(x_1, \ldots, x_D)$ as

$$z = \alpha + \beta_1 \log x_1 + \cdot\cdot\cdot + \beta_D \log x_D + \text{error} \tag{7}$$

where $\beta_1 + \cdot\cdot\cdot + \beta_D = 0$ is the logcontrast condition ensuring that we deal exclusively with logratios. For the Darss Sill example this is indeed so. With the amalgamated Darss Sill data set used by Martin-Fernandez, Barcelo-Vidal, and Pawlowsky-Glahn (1997) we find that the residual sum of squares of the logcontrast regression model is $2.4667 \times 10^{11}$ compared with $2.8452 \times 10^{11}$ for the raw component regression model, a reduction of 13.3%.

## Logratio Analysis of Subcompositional Hypotheses

Improved reliability of regression is by no means the only reason for the use of logcontrast models. As Woronow (1997b, p. 159) admits in his statement:

> (i) "The fact that deleting any single, arbitrary component yields the same quality regression is the good news. The bad news is that different values of the coefficients ensue when different variables are deleted. This does not imply that something is wrong with the method for arriving at a predictive equation. However, it makes clear that it is impossible to interpret the relative importance of variables from the magnitude or signs of their regression coefficients. . . . A corollary is that it is equally impossible to concoct a reliable geo-story for the values of these coefficients. Any such story lacks statistical basis, whether it makes use of the coefficients' raw magnitudes, their partial F-values or their beta coefficients."

In other words, raw regression can achieve nothing other than a prediction of sorts, not necessarily a reliable one. This brings into focus the relevance of the logcontrast type of regression. First it does allow investigation of the importance of the parts of the composition by allowing the investigation of subcompositional hypotheses. When we say that part $D$ of a $D$-part composition is unimportant we are really saying that the subcomposition consisting of the parts $1, \ldots, D - 1$ achieves the same explanation as the full composition. Even in the Darss Sill example with $D = 8$ we could, for example, ask if the last granulometric component is really contributing to the prediction. We can do this within model (7) by simply testing the hypothesis that $\beta_8 = 0$, involving a simple statistical F test. The residual sum of squares under the hypothesis is $2.4668 \times 10^{11}$, leading to an F value of 0.042 at (1,1274) degrees of freedom, clearly not significant. We can thus conclude that part 8 of the granulometric composition contributes nothing to the Easting prediction. If, however, we ask if the subcomposition consisting of parts $1, \ldots, 6$ is sufficient for predictive purposes, we then test the hypothesis that $\beta_7 = \beta_8 = 0$. The residual sum of squares under this hypothesis is $2.8869 \times 10^{11}$, giving a highly significant F-value of 108.5 at (2,1274) degrees of freedom. Hence we would conclude that component 7 cannot be dropped from the prediction process in addition to component 8.

This ability of the logcontrast regressor to explore the whole lattice of subcompositional hypotheses is in stark contrast to the failure of the linear regressor that essentially cannot deal with subcompositions because of its basic subcompositional incoherence. With the use of the logcontrast regressor, preserving ratios whatever subcompositional hypothesis is considered, we have an ideal mechanism for the study of the importance of the different parts of the composition. Moreover, as has been pointed out many times (Aitchison, 1983, 1986, 1997) logcontrasts have the ability to capture the often curved nature of compositional datasets, while also providing excellent approximations to linear configurations. This linear approximation results from the fact that the graph of the logarithmic function is almost linear over part of its range. To argue that the linear model could be extended to allow curvature by including quadratic terms would be using a sledge hammer to crack a predictive nut while at the same time confounding further any possibility of detecting the important predicting subcompositions.

What has been said above about regression applies equally well to discriminant analysis, where again in Woronow's (1997b) example of Erathem categorizing siltstones from their (CaO, MgO, FeO) compositions is discussed. No new issues arise in this dropping-one-out example and it seems pointless to investigate subcompositions in a situation where there is poor discrimination, 58.6% according to Woronow (1997b), a little better than coin-tossing assignment. For such discriminant analysis an excellent model—for example, for two categories—is the binary regression model with logcontrast predictor as in Eq. (7). For an example of the use of this, see Aitchison (1986, Section 12.6), where a whole lattice of subcompositional hypotheses is explored for discrimination between Permian and post-Permian rocks, with a 6-part subcomposition being found to be as successful as the full 10-part composition. An even more striking example is to be found in discriminating between two types of limestone from the Northern and Central Highlands of Scotland. Thomas and Aitchison (1997) show that of the 17-part (major-oxide, trace element) composition a simple major-oxide subcomposition (CaO, $Fe_2O_3$, MgO) provides excellent discrimination, equal to that of the full composition. Such a discovery that there is a simple and geologically interpretable explanation of the difference between the limestones can certainly be ascribed to logratio analysis and would not be discernible from the dropping-one-out techniques advocated in the Woronow (1997b) paper.

## DISCUSSION

The final paragraph of Woronow (1997b, p. 162) has the following conclusion:

(j) "Logratioing or other data transformations that decrease the number of independent variables also can remove the degeneracy. However, they do not facilitate interpreting the importance of a single compositional variable, and may complicate the matter further by

> sacrificing the simplicity of working with the natural units of the composition. Why go to unnecessary measures to accomplish the same end?''

We have seen in the constructive role that logratio analysis plays in addressing the whole range of compositional problems the fallacies in the above conclusion. First, the logratio transformation because of its one-to-one relationship with raw compositions in no way reduces the number of independent variables. Instead, by providing a sound, interpretable dependence structure for describing actual patterns of compositional variability, which allows the coherent investigation of subcompositional variability, the logratio transformation is admirably suited to investigating the importance or irrelevance of individual components. Instead of sacrificing the natural units of compositions, it in fact works explicitly with them. For it must surely be obvious that the essential nature of a composition is that relative, not absolute, magnitudes of components are the relevant ''units'' under study. It is these relative magnitudes or ratios that logratio analysis addresses, with the only role of the logarithm being the huge advantage in statistical tractability and interpretation that it brings. Logratios and logcontrasts also provide a simple and effective way of capturing the natural curvature that is often found in compositional data sets.

Above all, logratios and logcontrasts provide an excellent means of identifying or testing natural geological laws, such as those that involve equilibrium constants or the development of geological processes through time. Such exploration and testing has recently been reinforced by the development of models for geological processes in the form of differential perturbation processes (Aitchison, 1999; Aitchison and Thomas, 1998). In these the natural perturbation operator for describing compositional change (Aitchison, 1986, Section 2.8; 1997) forms the basis of a simple differential equation for describing the progress of a compositional process. The application of such models to compositional data sets for inference purposes leads inevitably to consideration of data in logratio and logcontrast form.

In all compositional data analysis, particularly in geology, the analyst should be aware that the observed compositions are often end products at various stages of some possibly long and unknown or poorly understood process. In such circumstances, although compositionial data may be unable to reveal the whole truth about the underlying process, they certainly form a substantial source of evidence. The validity of any hypothesis about the process should surely be converted into an equivalent hypothesis about these compositions and tested statistically against these compositional observations. On the other hand, if the compositional data is to be used to suggest possible hypotheses about the underlying process, then the nature of the variability among the compositions has to be suitably modeled and the consequent statistical analysis has to recognize the special nature of compositional data. In this second aspect, for example, the role of logcontrast principal component analysis may identify logratio-type laws similar to (6) which, in themselves or when converted into terms of concentrations, may give insights into the nature of the underlying process.

There have been a number of successful recent applications of logratio analysis in geology. The following selection gives an indication of the breadth of application. Anderson (1997) removes limitations of the Zn ratio in characterizing volcanic-hosted massive sulphide deposits by introducing a logratio version of the ratio. Cole and Drummond (1986), in a comprehensive study of precious metal ore deposits, investigate effects of various conditions on Ag and Au through the use of the Ag/Au logratio. Barcelo-Vidal, Pawlowsky-Glahn, and Grunsky (1997), Buccianti (1997), Buccianti, Vaselli, and Szabo (1997), Cardenas and others (1986), Grunsky and others (1992), Thomas (1997), and Zhou (1997) all use logratio analysis to resolve a variety of geological discrimination problems. None of these exploit the ability of logratio analysis to explore whether some subcomposition may achieve the same discriminatory power as the full composition. It might be of interest to investigate this subcompositional possibility along the lines of Thomas and Aitchison (1997). In geomorphologic studies, Ridenour and Giardano (1995a,b) use logratio analysis to identify the nature of hydraulic geometry. Renner (1991) and Weltje (1997) rely on a logratio analysis of residuals in assessing the success of endmember resolutions of compositional data.

As has been mentioned elsewhere, commitment to modeling patterns of variability of compositional data in terms of classes of distributions involving logratio covariance structures in no way limits their relevance to tackling hypotheses that are truly linear in character. We have given examples of these in Aitchison (1997) in relation to linear hypotheses and convex linear modelling in endmember analysis.

Denigrators of logratio techniques in compositional data analysis should perhaps reread the history of other transformations in statistical analysis. In particular, the logarithmic transformation, scoffed at by Karl Pearson and others with such questions as “What can be the meaning of the logarithm of a length?”, has become standard practice in most branches of science for particular data types, such as trace elements in geology, for example in the use of kriging techniques. Moreover, more exotic transformations, such as the Box–Cox transformation (Box and Cox, 1964), have now become standard tools in general linear modeling, even in geology (Barcelo, Pawlowsky, and Grunsky, 1996; Iyengar and Day, 1997).

Logratioing is a necessary measure for compositional data analysis and necessary measures are indeed required if the same uninterpretable ends as have been traditionally pursued by raw compositional data analysts over the last hundred years are to be replaced by sound scientific argument. Twice in Woronow (1997a, p. 99; 101) we are told that a prepared mind can readily interpret raw component analysis. It is surely reasonable to ask geologists to concentrate their prepared minds on relevant and reliable statistical inference.

## ACKNOWLEDGMENTS

I would like to thank Bruce Anderson and Eric Grunsky, reviewers of an earlier version of this paper, for their careful and constructive criticism.

## REFERENCES

Aitchison, J., 1981, A new approach to null correlations of proportions: Math. Geology, v. 13, no. 5, p. 175–189.

Aitchison, J., 1982, The statistical analysis of compositional data (with discussion): Jour. Roy Statist. Soc., v. B44, no. 2, p. 139–177.

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

Aitchison, J., 1986, The statistical analysis of compositional data: Chapman and Hall, London, 416 p.

Aitchison, J., 1997, The one hour course in compositional data analysis, or Compositional data analysis is easy, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG98, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 3–35.

Aitchison, J., 1999, Differential perturbation processes for compositional data analysis, in preparation.

Aitchison, J., and Thomas, C. W., 1998, Differential perturbation processes: a tool for the study of compositional processes, *in* Buccianti, A., Nardi G., and Potenza, R., eds., Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology: De Frede, Naples, p. 499–504.

Anderson, B., 1997, Potential problems in the characterisation of VHMS deposits using the Zn ratio, *in* V. Pawlowsky-Glahn, ed., Addendum to Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 1–10.

Allan, J. F., Satiza, R., Perefit, M. R., Fornart, D. J., and Sack, R. O., 1989, Petrology of lavas from the Lamont Seamont Chain and adjacent East Pacific Rise, 10° N: Jour. Petrology, v. 30, no. 5, p. 1245–1298.

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

Barcelo-Vidal, C., Pawlowsky-Glahn, V., and Grunsky, E. C., 1997, A critical approach to the Jensen diagram for the classification of a volcanic sequence, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 117–122.

Beard, J. S., and Day, H. W., 1988, Petrology and emplacement of reversely zoned gabbro-diorite plutons in the Smartville Complex, Northern California: Jour. Petrology, v. 29, no. 5, p. 965–995.

Box, G. E. P., and Cox, D. R., 1964, The analysis of transformations: Jour. Roy. Statist. Soc., v. B26, no. 2, p. 211–542.

Buccianti, A., 1997, Multivariate analysis to investigate Cl distribution in rocks from different settings: Math. Geology, v. 29, no. 3, p. 349–359.

Buccianti, A., Vaselli, G., and Szabo, Cs., 1997, Textural and chemical characterization of clinopyroxenes from ultramafic and granulaite xenoliths of the Carpathian-Pannonian region (Eastern Europe) by multivariate analysis, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 123–128.

Cardenas, A. A., Girty, G. H., Harison, A. D., Lahren, M. M., Knaack, C., and Johnson, D., 1986: Assessing differences in compositions between low metamorphic grade mudstones and high-grade schists using logratio techniques, Jour. Geology, v. 104, p. 279–293.

Chai, G., and Naldrett, A. J., 1992, The Jinchuan ultramafic intrusion: Cumulate of a high-Mg basaltic magma: Jour. Petrology, v. 33, no. 2, p. 277–303.

Chayes, F., 1949, On ratio correlation in petrography: Jour. Geology, v. 57, no. 3, p. 239–254.

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

Chayes, F., 1962, Numerical correlation and petrographic variation: Jour. Geology, v. 70, no. 4, p. 440–552.

Chayes, F., 1971, Ratio correlation: A manual for students of petrology and geochemistry: University of Chicago Press, Chicago, 99 p.

Chayes, F., and Kruskal, W., 1966, An approximate statistical test for correlation between proportions: Math. Geology, v. 74, no. 5, p. 692–702.

Cole, D. R., and Drummond, S. E., 1986, The effect of transport and boiling on Ag/Au ratios in hydrothermal solutions. A preliminary assessment and possible implications for the formation of epithermal precious-metal deposits: Jour. Geochemical Exploration, v. 25, no. 1, p. 45–79.

Davis, J. C., 1986, Statistics and data analysis in geology: Wiley, New York, 646 p.

Deer, W. A., Howie, R. A., and Zussman, J., 1982, Rock-forming minerals: Orthosilicates. Longman, London, 919 p.

Edwards, A. W. F., 1977, Foundations of mathematical genetics: Cambridge University Press, Cambridge, 119 p.

Eissen, J.-P., Juteau, T., Joron, J.-L., Dupre, B., Humler, E., and Al'Mukhameov, A., 1989, Petrology and geochemistry of basalts from the Red Sea Axial Drift at 18° North: Jour. Petrology, v. 30, no. 4, p. 791–839.

Fan, Q., and Hooper, P. R., 1989, The mineral chemistry of ultramafic xenoliths of Eastern China: Implications for upper mantle composition and the paleogeotherms: Jour. Petrology, v. 30, no. 5, p. 1117–1158.

Fan, Q., and Hooper, P. R., 1991. The Cenozoic basaltic rocks of Eastern China: Petrology and chemical composition: Jour. Petrology, v. 32, no. 4, p. 765–810.

Grunsky, E. C., Easton, R. M., Thurston, P. C., and Jensen, L. S., 1992, Characterization and statistical classification of Archean volcanic rocks of the Superior Province using major element geochemistry: Geology of Canada. Ontario Geological Survey Special Volume 4, p. 1397–1438.

Houthakker, H. S., 1960, Additive preferences: Econometrica, v. 2, p. 244–256.

Iyengar, M., and Day, D., 1997, Box-Cox transformations in Bayesian analysis of compositional data, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, Addendum, p. 40–47.

Kamenetsky, V. S., Sobolev, A. V., Joron, J. L., and Semet, M. P., 1995, Petrology and geochemistry of cretaceous ultramafic volcanics from Eastern Kamchatka: Jour. Petrology, v. 36, no. 3, p. 637–662.

Krauskopf, K. B., 1979, Introduction to geochemistry: McGraw-Hill, New York, 617 p.

Krumbein, C., 1962, Open and closed number systems: stratigraphic mapping: Bull. Amer. Assoc. Petrol. Geologists, v. 46, p. 322–337.

Le Maitre, R. W., 1982, Numerical petrography: Elsevier, Amsterdam, 281 p.

Martin-Fernandez, J. A., Barcelo-Vidal, C., and Pawlowsky-Glahn, V., 1997, Different classifications of the Darss-Sill data set based on mixture models for compositional data, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 151–156.

Mosimann, J. E., 1962, On the compound multinomial distribution, the multivariate $\beta$-distribution and correlations among proportions: Biometrika, v. 49, no. 1, p. 63–82.

Mosimann, J. E., 1963, On the compound negative binomial distribution and correlations among inversely sampled pollen counts: Biometrika, v. 50, no. 1, p. 47–54.

Pawlowsky, V., 1986, Räumliche Strukturanalyse and Schätzung ortsabhängiger Kompositionen mit Anwendungsbeispeilen aus der Geologie: unpublished dissertation, FB Geowissenschaften, Freie Universität Berlin, 120 p.

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 measurements of organs: Proc. Roy. Soc., v. 60, p. 489–498.

Renner, R. M., 1991, An examination of the use of logratio transformations for the testing of endmember hypotheses: Math. Geology, v. 23, no. 4, p. 549–562.

Reyment, R. A., and Savazzi, S., 1999, A primier of multivariate statistical analysis in geology: in press.

Ridenour, G. S., and Giardino, J. R., 1995a, Logratio linear modelling of hydraulic geometry using indices of flow resistance as covariates: Geomorphology, v. 14, p. 65–72.

Ridenour, G. S., and Giardino, J. R., 1995b, Discriminant function analysis of compositional data: an example from hydraulic geometry: Physical Geography, v. 15, no. 5, p. 481–492.

Rock, N. M. S., 1988, Numerical petrology: Springer-Verlag, Berlin, 427 p.

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

Thomas, C. W., 1997, Closure, log-ratios and the journeyman geologist: A plea, *in* V. Pawlowsky-Glahn, ed., Addendum to Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 34–39.

Thomas, C. W., and Aitchison, J., 1997, Applications of logratios to the statistical analysis of the geochemistry of metamorphosed limestones from the Northern and Central Highlands of Scotland: The case for the Appin Group correlations: British Geological Survey Technical Report WA/98/03, 20 p.

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

Woronow, A., 1987, A book review: The statistical analysis of compositional data by John Aitchison: Math. Geology, v. 19, no. 5, p. 579–581.

Woronow, A., 1997a, The elusive benefits of logratios, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitate Politècnica de Catalunya, Barcelona, p. 97–101.

Woronow, A., 1997b, Regression and discrimination analysis using raw compositional data—Is it really a problem?, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitate Politècnica de Catalunya, Barcelona, p. 157–162.

Woronow, A., 1997c, Calculating an equilibrium liquid-line of descent and determining a parental magma composition, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 129–132.

Woronow, A., and Love, K. M., 1990, Quantifying and testing differences among means of compositional data suites: Math. Geology, v. 22, no. 7, p. 837–852.

Zhou, D., 1997, Effect of logratio transformation on classifying of compositions, *in* V. Pawlowsky-Glahn, ed., Proceedings of IAMG97, The Third Annual Conference of the International Association for Mathematical Geology: Universitat Politècnica de Catalunya, Barcelona, p. 102–105.