An appraisal of an iterative construction of the endmembers controlling the composition of deep-sea manganese nodules from the Central Indian Ocean Basin

R M Renner[1,*], B N Nath[2] and G P Glasby[2]

[1]*School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, Wellington, New Zealand.*
[2]*National Institute of Oceanography, Dona Paula, Goa 403 004, India.*
Corresponding author. e-mail: ross.renner@msor.vuw.ac.nz

This paper describes an estimation of endmember compositions followed by the assessment of those results by log-ratio variance analysis. As an appraisal, it deals only with the first objective of an endmember analysis namely, to identify endmembers if they exist by estimating their compositions. Following the creation of the endmember estimates, the computation of an array of log-ratio variances was a key innovation in this type of study. Log-ratio variances revealed intrinsic linear associations between the dominant elements on each of the estimated endmember compositions, largely confirming the endmember analysis. The dataset under study contained the concentrations of 16 elements in 93 samples of deep-sea manganese nodules from the Central Indian Ocean Basin. Many previous analyses of these nodules were undertaken to assess the economic potential of the deposits. This study by contrast, quantified the inter-element associations that account for the nodule compositions. Four endmembers were identified. The elements loaded on each were: (1) Mn, Zn, Ni, Cu, Mn-rich, (2) Fe, Ti, P, Co, Fe-rich, (3) Si, Al, Na, K, clay minerals, (3) Mg, ultramafic material, possibly including Mn, Cr, V, Ca, Na. These latter elements were also detected by their log-ratio variances to be associated with Mg on the 4th endmember.

1. Introduction

In this paper, a sample will be a single geological specimen, not a statistical sample. An element will either be an element or its oxide. A *composition* will be a collection of measurements that sum to 100%, although in general the sum may have any constant value. Accordingly, a compositional dataset will be an $n \times p$ array in which the total concentrations of a fixed collection of p elements, in strict order, in each of n samples, sum to 100%, and are laid out in that strict order in n rows of the array, thus defining the columns. The number of endmembers will be denoted by k. A *component* of a composition will be a specific element, and an element will be *dominant* in an endmember if its concentration in that endmember is greater than its concentrations in the remaining endmembers. The term correlation will always mean the Pearson correlation.

The purpose of this paper is to describe a long-established geometrical construction procedure for estimating both the number of endmembers and their compositions from a compositional dataset,

Keywords. Compositional data; endmember analysis; log-ratio variance; deep-sea manganese nodules; Central Indian Ocean Basin.

together with a numerical procedure for assessing those estimates. Both of these procedures completely avoid the possibility of the type of spurious results that may be produced by procedures based on the correlation coefficient.

1.1 *Subcompositional coherence*

It has been well-documented for over a century (Pearson 1897) that there are difficulties associated with the statistical analysis of compositional data, such as percentages or ppm. Three of these difficulties are: (1) Such data do not follow multivariate normal distributions thus rendering standard parametric statistical tests and estimation procedures inappropriate and possibly misleading. (2) The negative bias property. That is, at least one covariance and hence at least one correlation, must be negative, therefore the correlations are prevented from ranging freely between -1 and $+1$. (3) Regression, covariance, and correlation coefficients between specific pairs of components vary unpredictably according to the presence or absence of other components in the compositions. Quantities, like regression and correlation coefficients that exhibit this unpredictable variability are said to lack *subcompositional coherence.* It follows from this property, that correlation coefficients in particular, that are formed from compositional data are not only not absolute, but also frequently spurious. Standard multivariate procedures based on them such as Principal Components Analysis and Factor Analysis are unreliable and worse, intrinsic or natural associations between elements inferred from strong positive correlations, have been shown to be potentially false (Renner 2012).

Since endmember analysis is fundamentally concerned with the intrinsic associations within suites of elements, two quite different approaches that are described in this paper to detect such associations are: (a) the estimation of endmember compositions, which avoids the issue of subcompositional incoherence altogether (Renner 1989) and, (b) the derivation of a statistic, the logratio variance (Aitchison 1986) that possesses subcompositional coherence and that does measure, the extent or otherwise of the intrinsic relationship between pairs of components (Appendix).

1.2 *The development of endmember analysis*

Many attempts to solve the endmember composition problem have relied on interpretations of the loadings on various factors computed by R-mode factor analysis, a statistical procedure based on the matrix of correlation coefficients. Two intractable difficulties with this approach are

that factor analysis is simply not a mathematical model for a geophysical mixing process and equally fundamentally, correlations formed from compositional data are not subcompositionally coherent.

The difficulties with analyzing compositional data were made known in Geosciences by the early works of Chayes (1960, 1962, 1983). A recognition by geoscientists of the problems described by Chayes, led to the development of endmember analysis. This was for the particular case that a compositional dataset could be the realization of a geophysical mixing process involving a small number k of fixed compositions. The problem was to estimate k and the compositions of the endmembers, with the strict condition that all the abundances (mixture coefficients) of the estimated endmembers in all the samples must be positive for a solution to be feasible. The early numerical algorithms for a solution to this problem that evolved over time were proposed by Imbrie and Van Andel (1964), Miesch (1976), Full *et al.* (1981), and Leinen and Pisias (1984). These proposals attempted to utilize Q-mode factor analysis which does indeed reveal the similarity of compositions as proximate points on a unit hypersphere, but is simply not a mathematical model for a geophysical mixture process, and cannot progress to feasible solutions. Chen and Owen (1989) advocated using linear programming methods. The appeal of this procedure is that it cannot construct negative endmember abundances but it will produce feasible solutions for any arbitrary, non-extreme endmember estimates.

The endmember analysis adopted for this study is a conservative technique, based on the procedures originally developed by Renner (1988, 1989, 1993, 1995, 1996). Geometrically, the samples of a compositional dataset are represented by a set of data-points whose coordinates are the concentrations of the individual elements in each of the samples. Endmembers, if they exist, must be extreme points whose compositions define the vertices of a simplex that completely encloses the data-points. Examples of a simplex are, a line interval for $k = 2$ endmembers with one vertex at each end, a triangle, a tetrahedron, or a pentatope, for $k = 3, 4$ or 5 endmembers, respectively. All the data-points must be interior to the simplex. When that is the case, the composition of each data-point is a simple mixture of the endmember compositions, the extreme points. The mixture coefficients will be all positive (or zero) that sum to 100%.

When the number of endmembers, k, is assumed known, the construction procedure executed in this study involves first identifying k, the most remote data-points in the dataset. That is, those with suites of elements with approximately the

largest and least concentrations of individual elements. The algorithm computes the mixture coefficients for all the samples as if these initial extremes were endmembers. When any of the mixture coefficients are negative, the algorithm moves incrementally outwards from the initial extremes to neighbouring points, in directions uniquely determined by the negative mixture coefficients. Hence it creates new extremes and computes a new set of mixture coefficients. This process is repeated until the mixture coefficients are all positive, ensuring a feasible solution. Then the constructed extremes are the vertices of the closest fitting simplex that completely encloses all the data-points and are therefore, the estimated endmembers. The construction procedure is iterative. It tends to retain the initial groupings of extreme concentrations, provided they lead to feasible solution. It also remains as close as possible to the data-points, thus ensuring viable, if not necessarily, the most extreme feasible solutions possible. Convex combinations (mixture coefficients summing to 100%) of those extremes, are required to account well for the compositions of all the samples in the dataset.

Weltje (1997) also developed an iterative procedure that was designed to converge to an 'optimal' solution for the endmember compositions. It relies on the provision of 'sensible' initial endmember compositions.

1.3 *The log-ratio variance (LRV)*

In a distinctly different procedure, the log-ratio variance (LRV) due to Aitchison (1986) measures the magnitudes of the departures from strictly proportional intrinsic (or natural) associations between pairs of elements. In the unlikely case that the relationship between two elements was strictly proportional (linear), that is a fixed constant for all the samples, the LRV between them would be zero (Appendix). However, given the variability of geochemical data, it is how close the LRV is to zero, that implies an intrinsic linear association. Like the correlation coefficient, the LRV is independent of the scales of measurement of the individual components. Its most fundamental difference is that unlike the correlation coefficient, it is subcompositionally coherent (Appendix). That is, the LRV between a pair of elements is invariant to the removal of the concentrations of one or more of the other elements from all the samples to form subcompositions. In this context, a subcomposition is the resultant sample data after the restoration of the 100% sum. Clearly, an intrinsic association between two elements is absolutely unrelated to the presence or absence of the concentrations of other elements in a composition. By way of a

simple example, if two elements exist in all the samples bonded together in a chemical compound, that chemical compound is a physical property of the samples. It is not an artefact of the total concentrations of the elements in the compositions or subcompositions, whatever else the two elements in prticular may be bonded with. (A subcompositional transform is equivalent to a change of scale in a row of the dataset; see Appendix.)

If a suite of elements belongs to a single endmember, then the pair-wise departures from linear associations of these elements as determined by the data should be relatively small. This is because the formation of mixtures does not alter the intrinsic or natural associations between the elements, and a mixture is itself after all, a linear combination. Thus, the LRV provides an independent set of measurements for assessing the validity of the initial endmember analysis. It can also reveal intrinsic associations between elements that may not be so evident in the resulting endmember estimates. This property is noted further on concerning the makeup of an endmember with ultramafic properties, and also the identity of a sample with ultramafic content which had been excluded from the analysis.

1.4 *Deep sea manganese nodules*

In 1987, India was allocated a mine site in the Central Indian Ocean Basin (CIOB). Since then, a substantial amount of data has been collected on nodules from the CIOB. Initially, the results of studies on the data were classified, but have now been released into the public domain. Deep-sea manganese nodules from the CIOB have been described in detail by Jauhari and Pattan (2000), Valsangkar (2001), Jauhari and Iyer (2008), Mukhopadhyay *et al.* (2008), Sarkar *et al.* (2008), and Vineesh *et al.* (2009). Jauhari and Pattan (2000) carried out an R-mode factor analysis of the composition of nodules from the CIOB and identified three factors representing hydrogenous, diagenetic, and aluminosilicate phases in the nodules.

Much of the work on CIOB nodules was undertaken principally to assess the economic potential of these deposits. In this paper results are reproduced of endmember and log-ratio variance analyses, each undertaken on a database comprising the concentrations of 16 elements in 92 of the 93 samples taken from the CIOB (one sample was excluded because it appeared to be a serious outlier). The analyses were conducted to independently identify the principal intrinsic combinations of minerals controlling nodule compositions in the region. The inclusion of *Loss on Ignition* (LOI) and *Residue* made a total of 18 concentrations for each sample.

2. Numerical methods

2.1 *Bilinear unmixing*

Let p be the number of element concentrations in each of n sample compositions of an $n \times p$ geochemical dataset. Then an endmember analysis is a procedure by which each sample composition may, within tolerable errors, be expressed as a mixture. That is, a linear combination summing to 100%, of a small number $k\,(k << p)$, of fixed extreme compositions, the endmember estimates. If these estimates are merged with the dataset making a total of $n + k$ compositions, then they will, between them, contain the largest and least concentrations of each element in the dataset. It is this property that identifies the element and/or mineral associations. Bilinear unmixing is used to describe the situation when all the parameters have to be estimated from the data (Menke 1984; Weltje 1997). These parameters include the number k of endmembers, the compositions of the endmembers, and the proportions (abundances or mixture coefficients) of the endmember estimates in each sample. This study describes only the estimates for k and the compositions of the endmembers, although the mixture coefficients are, indeed, also computed by the iterative endmember construction procedure. Accordingly, endmember estimates were obtained as the components of $k = 4$ extreme points (vertices) of a tetrahedron in $p = 18$-dimensional space that most closely enclosed the compositional data-points. Since all the estimates of the observed data-points were interior points of this tetrahedron, each of them could be expressed as a simple mixture (a convex combination) of the extreme points. Applications of this particular construction are described in detail by Renner *et al.* (1997, 1998).

2.2 *A note on statistical methods*

The construction procedure used here is not a statistical analysis. No underlying probability distributions are assumed for the element concentrations. Hence, no estimates for the extreme compositions are contained in confidence intervals, and there can be no tests of significance. Moreover, in this study a statistical analysis of the residuals (errors in estimation, Renner 1991, 1993) is not provided. Assumptions about the probability distributions of geochemical data are difficult to justify. In a special case, Aitchison and Bacon-Shone (1999) have described the pattern of variability of particular target compositions assuming the distributions were approximated by logistic normal and logistic skew normal probability distributions, and that certain information about the sources was

given. More recent Bayesian statistical approaches to the identification and distributions of endmembers have been described by Palmer and Douglas (2008). In particular, Dobigeon *et al.* (2009) and Zare and Gader (2011) describe studies into the application of Bayesian statistical methods of endmember analysis in the field of hyperspectral image processing, which is an area of research into pattern recognition.

Nonetheless, a basic property of the iterative construction method adopted for this study is that it makes no assumptions about the variability of the data. It is based entirely on the geometry of the data-points. Consequently, it completely bypasses the shortcomings of all multivariate procedures based on the correlation coefficient (e.g., Principal Components and Factor Analysis).

2.3 *The log-ratio variance and proportional relationships*

The LRV between two elements, X_1 and X_2, denoted by $\mathrm{LRV}\,(X_1/X_2)$, was introduced by Aitchison (1986). It is defined to be the sample variance of $\log\,(X_1/X_2)$ for corresponding pairs of elements, X_1 and X_2, in the dataset. It is shown in the Appendix that $\mathrm{LRV}\,(X_1/X_2) = \mathrm{LRV}\,(X_2/X_1)$, so that like the correlation matrix, the complete set of LRVs forms a symmetric array. This array differs from the correlation matrix in two essential respects. All the entries in the main diagonal are zeros rather than ones, and there is no limit on the magnitudes of all other entries except they must be ≥ 0. If the relationship between two elements were strictly proportional, that is $X_1/X_2 = m\,(> 0)$, a constant for all pairs of (X_1, X_2), then $\mathrm{LRV}\,(X_1/X_2)$ and $\mathrm{LRV}\,(X_2/X_1)$ would be zero (Appendix). However, given that geochemical data is notoriously noisy, the observation of such a strict proportionality would be extremely unlikely. So it is the proximity of a particular LRV to zero, in comparison with the LRVs of other pairs, that implies the level of proportionality, or systematic variation. Moreover, if a number of elements have mutually pair-wise small LRVs, then they should be associated together as a suite on an endmember since they would be jointly linearly related.

3. Results

An endmember analysis as described by Renner *et al.* (1997, 1998) was undertaken on the CIOB data, except that as an appraisal, it dealt only with the first objective of an endmember analysis namely, to identify endmembers if they exist

by estimating their compositions. It was followed by the computation of the Compositional Variation Array of LRVs (Appendix). Both procedures were based on the 92 compositions formed by merging two collections of deep-sea manganese nodules. These collections consisted of 34 and 58 samples respectively which, were sampled in the CIOB, between 9°–18°S and 72°–83°E, and analyzed for the same 16 elements. The 5th sample in the combined dataset, labeled SS2/92A, was excluded from the analyses. Its concentrations of Cr_2O_3 (0.038%), CaO (3.75%), and MgO (5.58%), were so high compared to all other samples, that its data-point was quite remote from the main body of the data. Such an apparent outlier may have been a perfectly valid though somewhat exceptional case. However, if it had been due to contamination or to measurement errors, its inclusion could have erroneously distorted the endmember construction process. Together with the 16 elements, LOI and *Residue* made a total of 18 components in each composition, all summing to 100%.

3.1 *Endmember estimates*

In order to provide equal weights to all elements, the concentrations of each one were divided throughout by its maximum concentration, forming a unit-range from zero to one for each element. The minor elements, usually measured in ppm, would otherwise appear as noise relative to the percentage data, the latter being on a scale 10,000 times larger in magnitude. Geometrically, without this transformation, the approximate dimensionality of the raw data points would in any case, tend to be biased towards the number of most abundant major elements. (It is an important property of the LRV that, like the correlation coefficient, it is invariant to any transformation such as this, which is equivalent to a change of scale for each element; see Appendix.)

A numerical procedure known as a Singular Value Decomposition (SVD), which orthogonally rotates the reference axes, was executed on the dataset of unit-ranges. It revealed that 99.13% of the total of the sum of squared coordinates (SSQ) on the rotated 18-dimensional reference axis system was accounted for by just 3 of its axes, whilst a fraction more, SSQ = 99.37%, was accounted for by 4 axes. For each of these 3-axes and 4-axes systems, the SVD computed data-points corresponding to, and closest to each of the of raw unit range data-points, converted back to compositional data points by the inverse of the unit-range transformation. In this way, two datasets of approximations to the dataset of raw compositions were created.

The quality of the approximations of the elements in these three and four dimensional datasets was assessed by the coefficient of determination, r^2. This is a measure of the proportion of the variation accounted for by the least squares line fitted to the points of pairs of observed and corresponding approximated concentrations of each element in a reduced axes system. Ideally, these points lie on a straight line through the origin with slope 1. When that is the case, $r^2 = 1\,(100\%)$. Hence, a value of $r^2 < 0.50\,(50\%)$ is clearly a poor fit since less than 50% of the variation is predicted by the line.

For just two components in the 3-axes system, $r^2 < 0.50\,(50\%)$ namely, for Cr_2O_3, $r^2 = 0.05\,(5\%)$, and for MgO, $r^2 = 0.29\,(29\%)$. However, in the case of the 4-axes system, $r^2 = 0.08\,(8\%)$ for Cr_2O_3, while for MgO, $r^2 = 0.82\,(82\%)$, a substantial rise from 29% (table 1), and a good fit. A five axes system did not raise r^2 above 0.50 (50%) for Cr_2O_3. Accordingly, since the SVD evidence indicated a single endmember to account for MgO, extreme compositions with the required properties of endmembers were to be sought in the 4-dimensional reference system. A second algorithm was then executed to iteratively construct the four endmember estimates from the four dimensional dataset. Given the very slight increase in the SSQ from 99.13% to 99.37%, there remained a geometrical limitation associated with the choice of the fourth dimension. All the data-points had to be very close to the plane of the triangle whose vertices were the other three extreme points, implying that the simplex would be a flattened tetrahedron. This made the convergence of the iterative procedure to a fourth extreme somewhat arbitrary.

The four required endmember estimates that were iteratively constructed are displayed in table 1. In that table, LOI and *Residue* are the 13th and 18th components respectively. In each sample composition, *Residue* is defined as 100% minus the total of the concentrations of all other components. It is necessary in this type of endmember analysis that all elements are measured on the same scale (% in this case) and all compositions sum to a fixed constant (again, 100% in this case).

(1) In table 1, the elements that are dominant on *Endmember I* are Mn, Zn, Ni, and Cu. This endmember is characterized by its large concentration of Mn, that is consistent with a Manganese phase (Jauhari and Pattan 2000).

(2) Elements that are dominant on *Endmember II* (table 1) are Fe, V, Ti, Ca, P, and Co. This endmember features a large concentration of Fe which is consistent with an Iron phase (Jauhari

Table 1. *Endmember analysis of the compositions in percentages of 92 manganese nodule samples for 16 elements plus LOI and Residue, together with the coefficients of determination (r^2). The largest values of each element are marked in bold.*

Element no.	Element/oxide	Endmember I	Endmember II	Endmember III	Endmember IV	r^2
1	**Fe_2O_3**	3.5564	**24.2572**	8.3008	10.6612	0.90
2	**MnO**	**51.6043**	24.4968	0.3764	34.2201	0.96
3	**Cr_2O_3**	0.0183	0.0224	0.0224	0.0210	0.08
4	**V_2O_5**	0.0645	**0.0879**	0.0139	0.0589	0.81
5	**TiO_2**	0.0000	**1.1203**	0.3865	0.5387	0.92
6	**CaO**	2.0863	**2.8003**	1.3156	2.1809	0.55
7	**K_2O**	0.1260	0.8254	**3.4580**	1.4669	0.90
8	**P_2O_5**	0.0609	**0.6976**	0.1267	0.4872	0.92
9	**SiO_2**	3.6409	14.9516	**59.1805**	14.3578	0.96
10	**Al_2O_3**	0.4655	4.7889	**14.1171**	7.6720	0.93
11	**MgO**	3.1603	2.0517	1.7431	**5.1913**	0.82
12	**Na_2O**	2.2275	2.0295	**3.9781**	1.7750	0.62
13	**LOI**	**18.7858**	16.4203	6.2276	16.1614	0.90
14	**Zn**	**0.2158**	0.0156	0.0183	0.1868	0.87
15	**Ni**	**2.2744**	0.6459	0.1468	0.6498	0.85
16	**Co**	0.0245	**0.2514**	0.0000	0.2266	0.85
17	**Cu**	**2.3034**	0.0020	0.0000	0.7421	0.94
18	**Residue**	**9.3851**	4.5354	0.5880	3.4023	0.69

and Pattan 2000). The presence of Ca and V is discussed further on.

(3) The elements that are dominant on *Endmember III* (table 1) are K, Si, Al, and Na. This endmember is therefore consistent with an Alumino-silicate phase (Jauhari and Pattan 2000).

(4) Only one element is dominant on *Endmember IV* (table 1) and that is Mg. It is associated with ultramafic material. This endmember also has the second largest concentration of Mn.

3.2 *Log-ratio variances*

The grouping of elements revealed by collections of pair-wise mutually small LRVs between the elements can be compared to the grouping of dominant elements in the endmembers obtained by an endmember analysis. LRVs that are all pair-wise close to zero within a group imply mutually linear intrinsic associations between the elements of such a group. This should be detected by an LRV analysis if the group is part of every mixture. Accordingly, a two-way compositional variation array setting out the LRVs as described in the Appendix was created by the compositional data software package, CoDaPack (Comas and Thió-Henestrosa 2011). Since the complete array of LRV values is symmetric (like a correlation matrix), only LRV values above the diagonal of that array are displayed in table 2 (as is the case with CoDaPack).

In the absence of a known probability distribution for the LRVs, CoDaPack determines the 5th, 25th, 75th and 95th percentiles to be the critical points of the values in the LRV array. It is to be noted that since they are sample statistics, the precision of the observed percentiles is dependent on the randomness of the sampling methods and the size of the statistical sample in this case is $n = 92$. It is assumed for this study that both these requirements are satisfied. The LRVs that are less than the 25th percentile are the closest to zero of all the values in the array, indicating the strongest associations between pairs of elements in the data. Values between the 25th and 75th percentiles imply weak associations, while those greater than the 75th percentile are measures of the weakest associations, being farthest from zero. These ranges of values are highlighted in the output from CoDaPack, and are correspondingly identified in tables 2 and 3(a–d). That is, since the 25th and 75th percentiles are 0.0580 and 0.1840 respectively; the LRV values in tables 2 and 3(a–d) are in italics for values < 0.0580, bold face for values > 0.1840. The values in between 0.0580 and 0.1840 are in plain type.

(1) Table 3(a) sets out the LRVs between Mn and mutually related elements. Manganese is dominant on *Endmember I* (table 1). It is evident from table 3(a) that pair-wise LRVs between each of Mn, Ni, Zn were ≤ 0.045, and between each of Cu, Ni were ≤ 0.052. However, the LRV(Mn/Cu) = 0.084, exceeding the 25th percentile. Nevertheless, this value is the smallest LRV between Cu and all elements other than Ni and Zn. Indeed, all other LRVs involving Cu

are greater than the 75th percentile (table 2) except for Mg, where LRV(Mn/Mg) = 0.133 (tables 2 and 3a). That is, Cu is weakly associated with Mg, but very weakly associated with all remaining elements other than Mn, Ni and Zn. Hence the LVR values for Mn, Ni, Zn, and Cu, evidently do identify these four elements as a suite, confirming their status in the endmember analysis as all dominant on *Endmember I* (table 1). These results are similar to the element associations reported by Jauhari and Pattan (2000).

The elements Mg, V, Ca are included in table 3(a) because of their relationship to Mn. However, the nine LRVs, between Mg, V, Ca and each of Ni, Zn, Cu all exceed the 25th percentile (table 3a) indicating weak pair-wise associations between the elements of these two suites. In table 3(d), the LRV evidence implies that Mn is a major component not only of *Endmember I* but also of *Endmember IV* which accounts for the relationships between Mn, Mg, V and Ca.

   Residue is dominant on *Endmember I* (table 1), and so its low LRV values with Mn, Ni and

weakly with Zn, are consistent with that result (table 3a). Nevertheless, its association with Mg, V and weakly with Ca are properties the LRVs also associated with *Endmember IV* (table 3d).

(2) Table 3(b) sets out the LRVs between Fe and mutually related elements. Fe is dominant on *Endmember II* (table 1). Evidence for an association between Fe, Ti and P was reported by Jauhari and Pattan (2000). Pair-wise LRVs between Fe, Ti, P, and also between P and Co, were 0.032 or less, indicating close associations. But LRV(Fe/Co) = 0.064 and LRV(Co/Ti) = 0.074 (table 3b) both of which are greater than 0.0580, the 25th percentile. However, the LRVs between Co and all other elements except V are greater than 0.074 (table 2). The four elements, Fe, Ti, P, Co, are dominant on *Endmember II* (table 1).

It is also the case that Ca and V are dominant on *Endmember II* (table 1). The LRV(Ca/V) = 0.018 (tables 3a and 3d), indicating a close association between Ca and V. However, the LRVs between these two elements and, Fe, Ti, P, Co were all

Table 2. *The log-ratio variance array. Variances < 0.0580 are in italics. Variances > 0.1840 are in bold.*

	MnO	Cr$_2$O$_3$	V$_2$O$_5$	TiO$_2$	CaO	K$_2$O	P$_2$O$_5$	SiO$_2$	Al$_2$O$_3$
Fe$_2$O$_3$	**0.2000**	0.0909	0.0701	*0.0172*	0.0807	0.1728	*0.0291*	0.1623	0.1323
MnO	0	0.0659	*0.0435*	**0.2588**	*0.0519*	**0.2105**	0.1767	**0.2274**	**0.1874**
Cr$_2$O$_3$		0	*0.0320*	0.1271	*0.0229*	0.0797	0.0958	0.0795	0.0602
V$_2$O$_5$			0	0.1065	*0.0181*	0.1486	0.0646	0.1540	0.1217
TiO$_2$				0	0.1092	**0.1881**	*0.0265*	0.1799	0.1439
CaO					0	0.1166	0.0660	0.1139	0.0853
K$_2$O						0	**0.1875**	*0.0176*	*0.0167*
P$_2$O$_5$							0	**0.1939**	0.1355
SiO$_2$								0	*0.0205*

	MgO	Na$_2$O	LOI	Zn	Ni	Co	Cu	Residue
Fe$_2$O$_2$	0.1546	0.1365	0.1097	0.3554	0.2646	0.0642	**0.4703**	**0.2045**
MnO	*0.0359*	0.0854	*0.0189*	*0.0452*	*0.0302*	0.1733	0.0841	*0.0228*
Cr$_2$O$_3$	*0.0401*	*0.0268*	*0.0226*	0.1329	0.0973	0.1184	**0.2190**	0.0748
V$_2$O$_5$	*0.0502*	0.0619	*0.0137*	0.1408	0.0924	0.0715	**0.2219**	*0.0551*
TiO$_2$	**0.2024**	0.1735	0.1525	**0.4299**	0.3436	0.0742	**0.5703**	0.2692
CaO	*0.0482*	*0.0372*	*0.0142*	0.1290	0.0914	0.0801	**0.2176**	0.0616
K$_2$O	0.1313	*0.0545*	0.1308	**0.2633**	**0.2427**	**0.2225**	**0.3807**	**0.2143**
P$_2$O$_5$	0.1388	0.1431	0.0973	**0.3246**	0.2623	*0.0319*	**0.4600**	0.1878
SiO$_2$	0.1562	*0.0459*	0.1385	**0.2942**	0.2490	0.2357	**0.3937**	**0.2266**
Al$_2$O$_3$	0.1027	*0.0449*	0.1047	**0.2531**	0.2255	0.1735	**0.3645**	**0.1900**
MgO	0	0.0625	*0.0240*	0.0729	0.0628	0.1456	0.1332	*0.0562*
Na$_2$O		0	*0.0402*	0.1419	0.1054	0.1655	**0.2172**	0.0911
LOI			0	0.0874	*0.0547*	0.1080	0.1521	*0.0334*
Zn				0	*0.0369*	**0.2989**	*0.0472*	0.0705
Ni					0	**0.2484**	*0.0521*	*0.0360*
Co						0	**0.4456**	0.1876
Cu							0	0.1087
Residue								0

Table 3 (a–d). *Log-ratio variances between the dominant elements Mn, Fe, Al, Mg, and their mutually related elements. Variances < 0.058 are in italics. Variances > 0.184 are in bold.*

Table 3(a). *Log-ratio variances between Mn and mutually related elements.*

	Mn	Ni	Mg	V	Zn	Ca	Cu	Residue
Mn	0	*0.030*	*0.036*	*0.044*	*0.045*	*0.052*	0.084	*0.023*
Ni		0	0.063	0.092	*0.037*	0.091	*0.052*	*0.036*
Mg			0	*0.050*	0.073	*0.048*	0.133	*0.056*
V				0	0.141	*0.018*	**0.222**	*0.055*
Zn					0	0.129	*0.047*	0.071
Ca						0	**0.218**	0.062
Cu							0	0.109

Table 3(b). *Log-ratio variances between Fe and mutually related elements.*

	Fe	Ti	P	Co
Fe	0	*0.017*	*0.029*	0.064
Ti		0	*0.027*	0.074
P			0	*0.032*

Table 3(c). *Log-ratio variances between Al and mutually related elements.*

	Al	K	Si	Na	Cr
Al	0	*0.017*	*0.021*	*0.045*	0.060
K		0	*0.018*	*0.055*	0.080
Si			0	*0.046*	0.080
Na				0	*0.027*

Table 3(d). *Log-ratio variances between Mg and mutually related elements.*

	Mg	Mn	Cr	V	Ca	Na	Residue
Mg	0	*0.036*	*0.040*	*0.050*	*0.048*	0.063	*0.056*
Mn		0	0.066	*0.044*	*0.052*	0.085	*0.023*
Cr			0	*0.032*	*0.023*	*0.027*	0.075
V				0	*0.018*	0.062	*0.055*
Ca					0	*0.037*	0.062
Na						0	0.091

greater than 0.065. Moreover, the LRV evidence implies that Ca and V belong to *Endmember IV* (table 3d).

(3) Table 3(c) sets out the LRV between Al and mutually related elements. Pair-wise LRVs between Al, K, Si, and Na were at most 0.055. These four are the dominant elements on *Endmember III* (table 1). Jauhari and Pattan (2000) found evidence of an association between Al, K, and P (partially), and separately between Al and Si.

Although unaccounted by the endmember analysis, Cr appears in both table 3(c) and table 3(d). In the former, it has an association with Na, LRV(Na/Cr) = 0.027.

(4) Table 3(d) sets out the LRVs between Mg and mutually related elements. Magnesium is dominant on *Endmember IV* (table 1) which is associated with ultramafic material. Manganese is high on this endmember which may account not only for its low LRVs with Mg in particular, LRV(Mn/Mg=0.036), but also with V and Ca (table 3d). Pair-wise LRVs between Mg, Mn, Cr, V, Ca are < 0.058, the 25th percentile, but LRV(Mg/Na) = 0.063. LRV(Mn/Na) = 0.085 and LRV(Mn/Cr) = 0.066. Although Cr was not accounted by the iterative four-endmember analysis (r^2 = 0.08, table 1), it appears here to be associated with Mg, V, Ca and Na. Its LRVs between these four elements are all ≤ 0.040.

These results show that LRVs, table 3(a–c), largely confirmed the constituents of the suites of dominant elements that were identified on the *Endmembers I, II* and *III*. In the case of *Endmember IV*, the LRVs not only isolated Mg from the other three endmembers (*viz.*, it was not associated with Ni, Zn, Cu in *Endmember I*), but also grouped it with Mn, Cr, V, Ca, and Na. Thus, the LRV evidence confirmed that *Endmember IV* is an ultramafic phase in the nodules studied here. In addition, the high concentrations of Mg, Ca and Cr in the 5th sample, SS2/92A, that led to its being omitted from both analyses, are probably not anomalies. In view of their mutually near zero LRV values in table 3(d), the high concentrations of Mg, Ca, and Cr in sample SS2/92A imply that the sample may be an extrapolation of the estimate for the ultramafic *Endmember IV*.

4. Conclusions

4.1 *Summary*

Four endmembers were identified by the endmember analysis. The elements found to be dominat on each were assessed by log-ratio variance analysis, and a summary of those findings follow:

- *Endmember I*: Dominant elements are Mn, Zn, Ni, and Cu. With an Mn content of more than 50%, it is clearly an Mn-rich endmember. The LRV analysis confirmed the grouping of these four elements. Although the association of Cu with Mn was somewhat weak, it was still stronger than all elements other than Mn, Zn and Ni.

The endmember analysis revealed that Mn is also high on *Endmember IV*, which would account for the LRVs that associated it with Mg, V and Ca on that endmember.

- *Endmember II*: Dominant elements are Fe, Ti, P and Co. Its high Fe content identifies it as an Fe-rich endmember. The LRVs between Co and each of Fe and Ti were rather weak, but stronger than with any other element except P. Accordingly, the LRVs confirmed the grouping of these four elements. Endmember analysis also estimated dominant concentrations for Ca and V on this endmember. However, although the LRV analysis did confirm that these two elements were indeed closely associated, it found that any associations between them and each of Fe, Ti, P, and Co were weak.
- *Endmember III*: Dominant elements are Si, Al, Na, and K. The collective dominant status of these elements was confirmed by the log-ratio (LRV) analysis. It is an endmember of alumino-silicate minerals.
- *Endmember IV*: Mg is the only dominant element that was estimated by the endmember analysis, although both Mn and Fe do have their second highest concentrations on this endmember. It represents ultramafic material derived from the mantle. The LRV analysis associated Mg with Mn, Cr, V and Ca, but the value of LRV(Mg/Fe) = 0.1546 is weak.

4.2 *Discussion*

The computation of an array of LRVs from a compositional dataset sets out distinct and essentially fundamental data on the relationships between pairs of elements. In particular, individual items of the array are a measure of the deviations from linearity in those relationships. The smaller the values of those deviations, the greater are the non-random or systematic associations between the elements. The existence of an intrinsic, geophysical relationship between two elements, such as a chemical bond or flow from a common source, is not conditional on the measurements of the concentrations of other elements that may or may not have been included in a compositional dataset. Unlike the correlation coefficient, the LRV is absolutely invariant to the inclusion or exclusion from a compositional dataset of the concentrations of other elements. Furthermore, it is also invariant to changes to the scales of measurement to either the individual components or the individual samples. It is an absolute constant between two elements for the particular dataset. Consequently, the mutual proximity to zero of the pair-wise LRVs of a group of elements, relative to their values for all other elements in the dataset,

jointly imply that for the given data, the group exists as a naturally occurring combination. This feature was utilized in this paper to assess the results of the numerical endmember analysis.

Accordingly, the value of the LRV analysis in an appraisal of this kind, is not only that it provides a robust means to assess the findings of the endmember analysis, but it also raises the possibility that it will bring fresh information to that analysis. In this study, it confirmed that Mn is a major component not only of *Endmember I* but also of *Endmember IV*. Furthermore, it showed that Ca and V were not dominant on *Endmember II*. They were indeed closely associated together, but with the components Mn, Mg, Cr and Na that it associated with *Endmember IV*. Given the immutable nature of the LRV, this implies that the inconsistency with the result of the endmember analysis is irrefutable, and that the result is an anomaly. The LRV analysis associated Na with alumino-silicates on *Endmember III*, as did the endmember analysis, it also associated Na with Ca and Cr on *Endmember IV*. Finally, the concentrations of MgO (5.58%), Cr_2O_3 (0.038%), and CaO (3.75%) in the excluded sample SS2/92A, are all greater than their corresponding concentrations on any of the estimated endmembers. As already noted, the LRV analysis associated these elements as a group with others on the ultramafic *Endmember IV*. It would appear then that sample SS2/92A is a denser form of this material, and possibly not the anomalous outlier that led to its exclusion.

There remains the issue of the reliability of procedures based on the correlation coefficient. Jauhari and Pattan (2000) reported that R-mode factor analysis was carried out on ferromanganese nodule data from the CIOB comprising major, trace and rare earth elements. Their analysis created three factors. Factor 1 had strong negative factor loadings of Fe, Ti, P (and, Y, Sr, and all the REEs). Factor 2 had strong factor loadings of Mn, Mg, Zn, Ni, Cu, Ca (and Mo, Ba). While Factor 3 had factor loadings of Al, K and P (partially). In addition, the crude correlation between Al and Si was 0.71. Although the database of that study was not the same as that analysed in this paper, there are evident correspondences between the results of the two studies. There is a reason for that. It can be shown that under quite mild restrictions, if two components are naturally associated (as in a chemical bond), then the correlation coefficient between them will be positive. That association is a sufficient condition for a positive correlation but crucially, it is not a necessary condition. That is, the converse is not true. A high positive correlation between two components is not proof of the presence of any natural association. This limitation and the errors it may entail directly affect an

R M Renner et al.

R-mode factor analysis. When the correlation matrix (less the approximations to the reduced variances in the diagonal) is 'factored', a unique solution for the array of factor loadings is not immediately obtained. There are indefinitely many such solutions. The particular solution that is commonly computed is that obtained by a 'Varimax' rotation of the factor vectors (Kaiser 1958, 1959). This rotational method maximizes the sum of the variances of the factor loadings by maximizing the spread between the largest and least loadings on each factor. For any rotated solution the factor loadings on the unobservable factors are the correlations between those factors and the observable components. Consequently in the case of a Varimax solution, the loadings of the most highly correlated components on a factor tend to be clustered, as do those of the least correlated components. In any event, whatever the rotational method, the crude correlation matrix between components complete with all its unknown spurious entries, directly determines the chosen loadings on the factors. So despite the similarity to the results of this paper, the classifications of the elements by Jauhari and Pattan (2000) are unsubstantiated.

In a final assessment, factor analysis is an extremely sophisticated mathematical/statistical model which bears no resemblance to the simple linear geophysical mixing model. In its most frequently employed form, the components are standardized random variables. The common factors are latent random variables which are also standardized and vary from sample to sample. The standard score of each element in a sample is the sum of the products of the fixed factor loadings and the variable associated factors, plus another variable factor that is unique to the component. Consequently the factors cannot possibly represent endmember compositions. The factor loadings which are themselves usually mislabeled 'factors' are the constant correlations between the components and the factor vectors as noted earlier. Neither the factor loadings nor their 'variances' usually expressed as percentages could in any way be related to endmember abundances. The factor analysis model is the complete opposite of an endmember model. In the latter, an element concentration in a sample is the variable mixture (endmember abundance) of the fixed concentrations of that element in the various endmembers.

If practitioners seek to detect endmembers from compositional data by merely identifying suites of mutually associated elements, then an inspection of the LRV array such as that output by CoDaPack is, for the time being, the most reliable method. In the future, more comprehensible transforms of the LRV array that mimic the correlation matrix should become available together with software for quickly identifying suites, and that will allocate an element to more than one suite when that possibility arises. The LRV analysis neither provides endmember compositions nor their abundances in each of the samples. That information requires an analysis of the geometry of the data-points such as that described in the earlier part of this paper.

Acknowledgements

Authors would like to thank Prof. J H Johnston, Victoria University of Wellington who confirmed the identity of the endmembers, Dr B N Nath, National Institute of Oceanography, Goa, India for supplying the primary data, and Dr Peter Beatty, Oncology Department, Whangarei Hospital, New Zealand, without whose clinical care, this paper would not have been written.

Appendix

A given geochemical dataset, denoted by $\{x_{ij}\}$, consists of the measurements on the concentrations of p elements in each of n samples (e.g., rock, sediment). That is, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, p$, so that x_{ij} is the concentration of the jth element in the ith sample. Accordingly, $\{x_{ij}\}$ represents an $n \times p$ rectangular array, with the observation vectors of the samples laid out along the rows, and the elements down the columns.

When the concentrations are all measured on the same scale, and, $x_{i1} + x_{i2} + \cdots + x_{ip} = \mathrm{C}$ (a fixed constant) for all i, then each row of the array $\{x_{ij}\}$ is a composition. Suppose $\mathrm{C} = 100\%$ and the pth (or any other) component is excluded from every sample; the remaining components are then rescaled to sum to 100%, each row of the $n \times (p-1)$ array $\{y_{ij}\}$, so formed, is a subcomposition. In the case that it is x_{ip} that is excluded for all i, the subcompositional component corresponding to x_{ij} is $y_{ij} = a_i x_{ij}$, where $a_i = 100/(100 - x_{ip})$ for each i. So the ratio of any two components of a subcomposition, y_{ij}/y_{ik}, is $a_i x_{ij}/a_i x_{ik} = x_{ij}/x_{ik}$, the same ratio as that of the corresponding components in the full composition. (Indeed, this result would be true for any other value of a_i.)

Abundance data is typically compositional, being expressed as percentages, ppm, etc. In order to overcome the well-documented difficulties associated with the statistical analysis of such data, Aitchison (1986) showed that transforming the data to log-ratios by forming typically, $v_{ij} = \log(x_{ij}/x_{ip})$, $j = 1, 2, \ldots, p-1$, it was possible in some cases to apply traditional multivariate statistical methodology to the $n \times (p-1)$ array $\{v_{ij}\}$.

In addition, Aitchison ([1986](#)) showed that given certain assumptions on the evolution of the data, $v_{ij} = \log(x_{ij}/x_{ip})$ is a component of a multivariate normal distribution.

A.1 Aitchison compositional variation array

If x_{ik} and x_{il} are the kth and lth concentrations in the ith sample, and $w_{ikl} = \log(x_{ik}/x_{il})$, then the log-ratio mean $\overline{w}_{kl}$, and log-ratio variance (LRV) s_{kl}^2, of the w_{ikl}, over all n samples, are given by:

$$\overline{w}_{kl} = \frac{1}{n} \sum_{i=1}^{n} w_{ikl} \qquad (1)$$

$$s_{kl}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (w_{ikl} - \overline{w}_{kl})^2 \qquad (2)$$

for $k, l = 1, 2, \ldots, p$, in equations ([1](#)) and ([2](#)). The square $p \times p$ array $\{\overline{w}_{kl}\}$ is anti-symmetric since $w_{ikl} = -\log(x_{il}/x_{ik}) = -w_{ilk}$ for all i, and so $\overline{w}_{kl} = -\overline{w}_{lk}$. The diagonal entries of $\{\overline{w}_{kl}\}$ are all zero. Then, $l = k$, $x_{ik}/x_{ik} = 1$, and $\log(1) = 0$. Similarly, the diagonal entries of $p \times p$ array $\{s_{kl}^2\}$ are also zero. There is no limit on the magnitudes of the non-diagonal entries of $\{s_{kl}^2\}$ except they must be ≥ 0. Hence $\{s_{kl}^2\}$ is clearly symmetric. Aitchison ([1986](#)) defined the Compositional Variation Array to be that $p \times p$ array which contains the values of $\{s_{kl}^2\}$ above its diagonal, and the values of $\{\overline{w}_{kl}\}$ below. (The diagonal entries are left blank.) This array is computed by the free, compositional data processing software package, CoDaPack (Comas and Thió-Henestrosa [2011](#)).

It is evident that the mean and variance of the log-ratios of two elements in any subcomposition are equal to those for the same elements in the full composition since the corresponding ratios are equal. This is a consequence of scaling all the elements of the ith row of $\{y_{ij}\}$ by an ith constant a_i. There is a similar result obtained by scaling the columns of $\{x_{ij}\}$ (or $\{y_{ij}\}$) that applies to just the LRV, s_{kl}^2. Suppose all the concentrations of $\{x_{ij}\}$ are percentages, then provided $b_k = 10,000$, the term $b_k x_{ik}$ is in ppm for all i. That is, the kth column of $\{x_{ij}\}$ is now scaled by the constant b_k.

Let $u_{ikl} = \log(b_k x_{ik}/x_{il})$. Then $u_{ikl} = \log(b_k) + \log(x_{ik}/x_{il})$, that is, $u_{ikl} = \log(b_k) + w_{ikl}$. So the mean $\overline{u}_{kl}$ of the u_{ikl} is given by:

$$\overline{u}_{kl} = \frac{1}{n} \sum_{i=1}^{n} (\log(b_k) + w_{ikl}) = \log(b_k) + \overline{w}_{kl}. \quad (3)$$

Hence, in the expression for the LRV, t_{kl}^2 of the u_{ikl}, the deviation from the mean is $(u_{ikl} - \overline{u}_{kl}) =$

$(w_{ikl} - \overline{w}_{kl})$, since $\log(b_k)$ cancels, so that t_{kl}^2 becomes:

$$t_{kl}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (u_{ikl} - \overline{u}_{kl})^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (w_{ikl} - \overline{w}_{kj})^2 = s_{kl}^2 \qquad (4)$$

which is the variance of the w_{ikl}. Summing up, the LRV between two elements is invariant for a composition and its subcompositions (i.e., two row transformations). This property is called subcompositional coherence. It follows that the LRV remains invariant even in the special case where two-part subcompositions are created of the form $X_1 + X_2 = 100\%$, by the reduction of all samples to just the two elements, X_1, X_2. By contrast, the correlation coefficient in that case between X_1 and X_2, would necessarily be -1 (or possibly undefined), irrespective of the relationship between X_1 and X_2. The LRV is also invariant to differences in the scales of the measurements of the elements (column transformations). In particular, they need not be on the same scale. That also happens to be a property of the (Pearson) correlation coefficient. The fundamental difference being, the LRV is immutable to both row and column transformations. Such results would seem to imply that it measures a basic underlying relationship between any two elements of a given dataset.

A.2 Intrinsic linear associations

The identification of intrinsic associations between the elements of geochemical datasets is vital to understanding the makeup of minerals, sediments, and other geological substances. A strong intrinsic association between Al and Si, for example, would indicate the presence of alumino-silicates. Traditionally, the Pearson correlation coefficient has been widely misused in the geosciences as a measure of linear association. Although it has been well-documented for over a century that the correlation coefficients between elements of a compositional dataset are spurious, and therefore, may or may not have any geochemical relevance (see Pearson [1897](#); Aitchison [1986](#); Pawlowsky-Glahn and Egozcue [2006](#)). Moreover, unlike the LRV, the correlation coefficient lacks subcompositional coherence. In fact, the correlation between a corresponding pair of elements of a composition and its subcompositions are not only not necessarily equal, but may actually contradict each other, one being a large positive, the other an equally large negative (Renner [2012](#)).

The potentially useful property of the LRV, s_{kl}^2, is that if there were a perfect linear relation between

two elements that is, $x_{ik}/x_{il} = m \, (> 0)$, a fixed constant of proportionality for all i, then $s_{kl}^2 = 0$. In this elementary situation, $w_{ikl} = \log(x_{ik}/x_{il}) = \log(m)$. Hence $\overline{w}_{kl} = \log(m)$, so $(w_{ikl} - \overline{w}_{kl})^2 = 0$ for all i, and hence $s_{kl}^2 = 0$. Geochemical data is of course, notoriously noisy. Even if two elements were bound up in all the samples in such a way as Al and Si are in an alumino-silicate, their ratios (e.g., x_{iAl}/x_{iSi}), would not be precisely constant. Moreover, the presence of these elements from more than one source mineral increases the variability in the log-ratio. Nonetheless, in the expression for s_{kl}^2, all the terms $(w_{ikl} - \overline{w}_{kl})^2 \geq 0$, so a value of s_{kl}^2 close to zero would imply all such terms are close to zero, and hence the kth and lth elements of the dataset would have an approximately linear, intrinsic association. When a subset of the elements of the dataset all have mutually pairwise LRVs close to zero, then their joint intrinsic associations would indicate that the elements of that subset form a suite such as the group of dominant elements in an endmember. The software package CoDaPack (Comas and Thió-Henestrosa 2011) outputs the Compositional Variation Array, with the lowest values of the LRVs highlighted (in blue), thereby expediting the identification of related elements.

A.3 An alternative correlation matrix

The array of log-ratio variances can be transformed into an array with the same symmetric layout as the correlation matrix except all entries would be non-negative but still ≤ 1. For example, $\exp(-\mathrm{LRV})$ and $\exp(-\sqrt{\mathrm{LRV}})$ are equal to one for $\mathrm{LRV} = 0$, and they both tend to zero with increasing LRV. In either case, a value of one signifies a perfect linear association, accounting for ones in the diagonal of the array, while a value of zero would signify an absence of linear association. An unresolved problem is to determine the statistical properties of any such transformation. In the case of the Pearson correlation coefficient, the statistical properties are known when a bivariate pair is assumed to be bivariate normal. Ironically, this is never the case for compositional data, despite the widespread citations of p-values obtained from the correlation matrices output by computer packages. It is clear that the properties of the LRV described here apply to all multivariate data, and not just compositional data. Hence empirical studies involving specified statistical distributions could compile for comparison, the results of Principal Components, and other multivariate analyses based on both correlation and LRV matrices. So, until the theoretical issues are resolved, such empirical studies could provide information on the variable behaviour of the LRV under differing initial conditions.

References

Aitchison J 1986 *The Statistical Analysis of Compositional Data*; Chapman and Hall, London.

Aitchison J and Bacon-Shone J 1999 Convex linear combinations of compositions; *Biometrika* **86** 351–364.

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. Geol.* **70** 440–552.

Chayes F 1983 Detecting nonrandom associations between proportions by tests of remaining space variables; *Math. Geol.* **15** 197–206.

Chen J C and Owen R M 1989 The hydrothermal component in ferromanganese nodules from the southeast Pacific Ocean; *Geochim. Cosmochim. Acta* **53** 1299–1305.

Comas M and Thió-Henestrosa S 2011 CoDaPack 2.0: A stand-alone, multi-platform compositional software; In: *Proceedings of the 4th International Workshop on Compositional Data Analysis* (eds) Egozcue J J, Tolosana-Delgado R and Ortego M I, pp. 1–10.

Dobigeon N, Moussaoui N, Coulon M, Tourneret J and Hero A O 2009 Signal processing; *IEEE Trans. Signal Process.* **57** 4355–4368.

Full W E, Ehrlich R and Klovan J E 1981 Extended Qmodel – objective definition of external endmembers in the analysis of mixtures; *Math. Geol.* **13** 331–344.

Imbrie J and Van Andel T H 1964 Vector analysis of heavy mineral data; *Geo. Soc. Am. Bull.* **76** 1131–1156.

Jauhari P and Pattan J N 2000 Ferromanganese nodules from the Central Indian Basin; In: *Handbook of Marine Mineral Deposits* (ed.) Cronan D S (Boca Raton: CRC Press), pp. 171–195.

Jauhari P and Iyer S D 2008 A comprehensive view of manganese nodules and volcanics of the Central Indian Ocean Basin; *Mar. Georesour. Geotechnol.* **26** 231–258.

Kaiser H F 1958 The varimax criterion for analytic rotation in factor analysis; *Psychometrika* **23** 187–200.

Kaiser H F 1959 Computer program for varimax rotation in factor analysis; *Educ. Psychol. Meas.* **19** 413–420.

Leinen M and Pisias N 1984 An objective technique for determining end-member compositions and for partitioning sediments according to their sources; *Geochim. Cosmochim. Acta* **48** 47–62.

Menke W 1984 *Geophysical data analysis: Discrete inverse theory*; Academic Press, Orlando.

Miesch A T 1976 Q-mode factor analysis of compositional data; *Comput. Geosci.* **1** 147–159.

Mukhopadhyay R, Ghosh A K and Iyer S D 2008 *The Indian Ocean Nodule Field: Geology and Resource Potential*; Elsevier, Amsterdam.

Palmer M J and Douglas G B 2008 A Bayesian statistical model for endmember analysis of sediment geochemistry, incorporating spatial dependencies; *Appl. Stat. J. Roy. St. Series C* **57** 313–327.

Pawlowsky-Glahn V and Egozcue J J 2006 Compositional data and their analysis: An introduction; In: *Compositional Data Analysis in the Geosciences: From Theory to Practice* (eds) Buccianti A, Mateu-Figueras G and Pawlowsky-Glahn V; *Geol. Soc. London, Spec. Publ.* **264** 1–10.

Pearson K 1897 Mathematical contributions to the theory of evolution. On a form of spurious correlation which may

arise when indices are used in the measurement of organs; *Proc. Roy. Soc. London* **60** 489–498.

Renner R M 1988 On the resolution of compositional datasets into convex combinations of extreme vectors; *Technical Report 88/02*; Institute of Statistics and Operations Research, Victoria University of Wellington, Wellington, New Zealand.

Renner R M 1989 On the resolution of compositional datasets into convex combinations of extreme vectors; Ph.D. Thesis, Victoria University of Wellington, Wellington, New Zealand.

Renner R M 1991 An examination of the use of the logratio transformation for the testing of endmember hypotheses; *Math. Geol.* **23** 549–563.

Renner R M 1993 The resolution of a compositional dataset into mixtures of fixed source compositions; *Appl. Stat. J. Roy. St. Series C* **42** 615–631.

Renner R M 1995 The construction of extreme compositions; *Math. Geol.* **27** 485–497.

Renner R M 1996 An algorithm for computing extreme compositions; *Comput. Geosci.* **22** 15–25.

Renner R M 2012 Statistical comparisons of heavy metal pollutants between seven regions of the Polish exclusive economic zone; *Environ. Earth Sci.* **67** 987–997.

Renner R M, Glasby G P and Walter P 1997 Endmember analysis of metalliferous sediments from the Galapagos Rift and East Pacific Rise between 2°N and 42°S; *Appl. Geochem.* **12** 383–395.

Renner R M, Glasby G P and Szefer P 1998 Endmember analysis of heavy-metal pollution in surficial sediments from the Gulf of Gdansk and the southern Baltic Sea off Poland; *Appl. Geochem.* **13** 313–318.

Sarkar C, Iyer S D and Hazra S 2008 Inter-relationship between nuclei and gross characteristics of manganese nodules, Central Indian Ocean Basin; *Mar. Georesour. Geotechnol.* **26** 259–289.

Valsangkar A B 2001 Mineral resources; In: The Indian Ocean–A Perspective, Volume 2, Oxford & IBH Publishing Co. Pvt. Ltd, New Delhi, pp. 585–643.

Vineesh T C, Nath B C, Banerjee R, Jaisankar S and Lekshmi V 2009 Manganese nodule morphology as indicators for oceanic processes in the Central Indian Basin; *Int. Geol. Rev.* **51** 27–44.

Weltje G J 1997 End-member modeling of compositional data: Numerical-statistical algorithms for solving the explicit mixing problem; *Math. Geol.* **29** 503–549.

Zare A and Gader P 2011 An investigation of likelihoods and priors for Bayesian endmember estimation; *Am. Inst. Physics Conf. Proc.* **1305** 311–318.