

# Statistical models for point-counting data

Pieter Vermeesch

Department of Earth Sciences

University College London

`p.vermeesch@ucl.ac.uk`

August 4, 2018

## Abstract

Point-counting data are a mainstay of petrography, micropalaeontology and palynology. Conventional statistical analysis of such data is fraught with problems. Commonly used statistics such as the arithmetic mean and standard deviation may produce nonsensical results when applied to point-counting data. This paper makes the case that point-counts represent a distinct class of data that requires different treatment. Point-counts are affected by a combination of (1) true compositional variability and (2) multinomial counting uncertainties. The relative significance of these two sources of dispersion can be quantified by a chi-square statistic and  $\chi^2$ -test. For datasets that pass the chi-square test for homogeneity, the ‘pooled’ composition is shown to represent the optimal estimate for the underlying population. It is obtained by simply adding together the counts of all samples and normalising the resulting values to unity. However, more often than not, point-counting datasets fail the chi-square test. The overdispersion of such datasets can be captured by a random effects model that combines a logistic normal population with the usual multinomial counting uncertainties. This gives rise to the concept of a ‘central’ composition as a more appropriate way to average overdispersed data. Two- or three-component datasets can be displayed on radial plots and ternary diagrams, respectively. Higher dimensional datasets may be visualised and interpreted by Correspondence Analysis (CA). This is a multivariate ordination technique that is similar in purpose to Principal Component Analysis (PCA). CA and PCA are both shown to be special cases of Multidimensional Scaling (MDS). Generalising this insight to multiple datasets allows point-counting data to be combined with other data types such as chemical compositions by means of 3-way MDS. All the techniques introduced in this paper have been implemented in the `provenance` R-package, which is available from <http://provenance.london-geochron.com>.

# 1 Introduction

The mineralogical composition of siliclastic sediments can be determined by tallying the occurrence of various minerals in a representative sample of (200-400, say) grains (Dryden, 1931; Van der Plas and Tobi, 1965; Weltje, 2002). Similarly, the fossil content of a deep sea sediment core may be characterised by tabulating the relative abundances of various species among >100 randomly selected specimens (Patterson and Fishbein, 1989; Buzas, 1990; Fatela and Taborda, 2002). Or palaeobiological environments may be reconstructed by tabulating the relative frequency of different types of pollen in a palaeosol or charcoal (Barkley, 1934; Clark, 1982; Weng et al., 2006).

These are all examples of multivariate counting experiments, in which the unknown proportions of different species of minerals, fossils or pollen are estimated by counting a finite number of randomly selected items from a representative sample. Despite the widespread use of this type of data in the Earth Sciences and related fields, their statistical analysis is demonstrably underdeveloped.

For example, there currently exists no agreed method to average multi-sample point-counting datasets, or to quantify point-counting data dispersion. Traditionally, these operations were done by taking the arithmetic mean and standard deviation, respectively. Unfortunately, this may easily produce nonsensical results. For example, Weltje (2002) shows that the common practice of using ‘2-sigma’ confidence bounds around the arithmetic mean can produce physically impossible negative values when applied to petrographic point-counts.

To solve these problems, Weltje (2002) argues that point-counts should be treated as *compositional* data, which are defined as “vectors representing parts of a whole that only carry relative information” (Pawlowsky-Glahn and Buccianti, 2011). According to this definition, compositional data can be renormalised to a constant sum (e.g., 100% if the composition is expressed as percentages, or 1 if fractions are used) without loss of information. Aitchison (1982, 1986) shows that the statistical analysis of such data is best carried out using a simple logratio transformation.

To illustrate this approach, let  $\{a_i, b_i, c_i\}$  be a three-component dataset, where  $a_i + b_i + c_i = 1$  for  $1 \leq i \leq m$ . Then this dataset can be mapped to a bivariate Euclidean data space as follows:

$$u_i = \ln(a_i/c_i) \text{ and } v_i = \ln(b_i/c_i) \tag{1}$$

53 After performing the desired statistical analysis (such as calculating averages and confidence regions) on  
 54 the transformed data  $\{u_i, v_i\}$ , the results can be mapped back to the ternary diagram by means of an inverse  
 55 logratio transformation:

$$\begin{aligned}
 a_i &= \frac{\exp[u_i]}{\exp[u_i] + \exp[v_i] + 1}, \\
 b_i &= \frac{\exp[v_i]}{\exp[u_i] + \exp[v_i] + 1}, \text{ and} \\
 c_i &= \frac{1}{\exp[u_i] + \exp[v_i] + 1}
 \end{aligned}
 \tag{2}$$

56 This procedure yields geologically meaningful (geometric) means and confidence regions. Weltje (2002)'s  
 57 adoption of logratio statistics to point-counting data represents a huge improvement over the 'crude' statistics  
 58 employed previously. But it does not solve all our problems. There are two crucial differences between point  
 59 counts and the classical compositional data discussed by Aitchison (1982, 1986).

60  
 61 First, point-counting data are associated with significant (counting) uncertainties, which are ignored by  
 62 classical compositional data analysis. For a single sample, this uncertainty is adequately described by multi-  
 63 nomial counting statistics (Section 6 of Weltje, 2002). But for larger datasets comprised of multiple samples,  
 64 existing procedures to construct confidence regions (as discussed in Section 7 of Weltje, 2002) are inadequate  
 65 because they lump together the 'observational' dispersion caused by counting statistics and the true 'geological'  
 66 dispersion. Bloemsma and Weltje (2015) describe a method to disentangle these two sources of uncertainty  
 67 in a logratio context. They show that deconvolution of (spectroscopic) count data into a scale vector and a  
 68 proportions matrix significantly improves multivariate analysis.

69  
 70 Second, point-counting data often contain zero values, which are incompatible with the log-ratio transforma-  
 71 tion defined in Equation 1. This problem also applies to the aforementioned approach by Bloemsma and Weltje  
 72 (2015). These authors circumvented the occurrence of sporadic zeros by replacing them with small positive  
 73 numbers. This and alternative 'imputation' strategies are further discussed by Martín-Fernández et al. (2003).  
 74 When the number of zeros is small, imputation is considered to have a minimal influence on the data covariance  
 75 structure. However, some point-counting datasets are dominated by zeros. So the presence of such values is not  
 76 a cosmetic problem, but a fundamental characteristic of this particular data type. The statistical treatment of  
 77 point-counting data needs to address this issue at a deeper level.

78

79 The present paper solves these long standing problems using established statistical methods adopted from  
80 other disciplines. Much of the paper is based on the work of Galbraith (2005) in fission track geochronology.  
81 The fission track method is based on the ratio of the number of spontaneous  $^{238}\text{U}$ -tracks to the number of  
82 neutron-induced  $^{235}\text{U}$ -tracks per unit area in accessory minerals such as apatite or zircon. This is equivalent  
83 to a simple two-component point-counting problem. Section 2 uses this equivalence to derive the concept of  
84 a ‘pooled composition’. We will show that the latter represents the most reliable (in terms of accuracy and  
85 precision) average of homogeneous point-counting data.

86

87 The analytical uncertainty of individual point-counting proportions may greatly vary between samples. Sec-  
88 tion 3 introduces Galbraith (1988)’s radial plot as a graphical means of visualising such ‘heteroscedastic’ data.  
89 Originally developed for fission track data, the radial plot can also be used to display point-counting ratios,  
90 which frequently occur in the Earth Sciences. Radial plots allow a visual assessment of the degree to which  
91 counting uncertainties can explain the observed scatter between multiple ratio estimates. Section 4 presents a  
92 formal statistical test to make this assessment more quantitative.

93

94 The pooled composition is only applicable to samples that pass this chi-square test for sample homogeneity.  
95 Multi-sample datasets that fail the chi-square test are said to be ‘overdispersed’ with respect to the counting un-  
96 certainties. The degree of overdispersion may be quantified by means of a continuous mixture model (Section 5).  
97 This model leads to the concept of a ‘central composition’ as a better alternative to the pooled composition of  
98 Section 2. Section 6 generalises the continuous mixture model from two to three (or more) components.

99

100 Finally, Section 7 introduces Correspondence Analysis (CA) as a useful ordination technique for multivariate  
101 point-counting data. CA is closely related to compositional Principal Component Analysis (PCA). But unlike  
102 the latter method, it does not suffer from the zero counts problem.

103

104 All the techniques discussed above will be illustrated with a combination of synthetic and real examples.  
105 The methods of Sections 2-6 will use the two datasets shown in Table 1. Data 1 consists of 20 random samples  
106 of 23–94 items each, which were drawn from a discrete trinomial distribution with 45% of component  $a$ , 45% of  
107 component  $b$  and 10% of component  $c$ . Data 2 comprises a further 20 samples that were drawn from a continuous  
108 distribution whose mode is the same as that of Data 1, but which adds 100% of dispersion around this mode.

109 Thus, Data 2 has two sources of dispersion (counting error and true population dispersion), whereas Data 1  
 110 only has one (counting error). Note that both datasets contain fewer counts per sample than is customary in  
 111 real world applications. But they are nevertheless realistic if we consider them to be ternary subcompositions  
 112 of higher dimensional datasets.

113

Data 1					Data 2				
#	<i>a</i>	<i>b</i>	<i>c</i>	<i>R</i>	#	<i>a</i>	<i>b</i>	<i>c</i>	<i>R</i>
1	16	18	4	38	1	23	24	5	52
2	25	17	3	45	2	60	24	7	91
3	18	18	0	36	3	45	43	12	100
4	7	14	3	24	4	2	53	4	59
5	12	10	3	25	5	8	32	10	50
6	32	30	13	75	6	53	21	23	97
7	35	38	13	86	7	1	6	3	10
8	20	20	7	47	8	2	17	1	20
9	10	9	3	22	9	10	10	4	24
10	29	36	5	70	10	2	35	3	40
11	34	34	9	77	11	29	21	3	53
12	22	47	12	81	12	2	13	0	15
13	9	9	2	20	13	3	9	0	12
14	37	36	13	86	14	34	1	0	35
15	46	25	16	87	15	28	19	4	51
16	50	37	7	94	16	49	11	3	63
17	28	34	8	70	17	0	72	2	74
18	39	50	6	95	18	55	28	13	96
19	44	36	10	90	19	7	8	3	18
20	28	21	4	53	20	20	5	2	27
<i>C</i>	541	539	142	<i>N</i> = 1222	<i>C</i>	433	452	90	<i>N</i> = 987

Table 1: Two synthetic ternary point-counting datasets. Data 1 was drawn from a single multinomial distribution with population proportions of 45%, 45% and 10% for components *a*, *b* and *c*, respectively. Data 2 was drawn from a continuous mixture of multinomial distributions whose true proportions were drawn from a bivariate logistic normal distribution with a geometric mean of 45% for *a* and *b*, 10% for *c*, and 100% dispersion with a correlation coefficient of -0.5 between the two logratio dimensions. *R*, *C* and *N* refer to the row, column, and total sums, respectively.

## 114 2 The pooled composition

115 Let us partition a population of objects (such as minerals, fossils or pollen) into two classes. Let  $\theta$  be the true  
 116 fraction of the first class, and  $(1 - \theta)$  the true fraction of the second class, respectively. Suppose that we have  
 117 collected  $m$  representative samples of randomly selected items from this population. For each of these samples,  
 118 let  $n_{i1}$  represent the number of items belonging to the first class and  $n_{i2}$  the number of items belonging to

119 the second class (for  $1 \leq i \leq m$ ). Further let  $n_{i\bullet}$  be the total number of counts in the  $i^{\text{th}}$  sample so that  
 120  $n_{i\bullet} \equiv n_{i1} + n_{i2}$ . The probability of observing  $n_{i1}$  and  $n_{i2}$  given  $n_{i\bullet}$  then follows a binomial distribution:

$$p(n_{i1}, n_{i2} | n_{i\bullet}) = \binom{n_{i\bullet}}{n_{i1}} \theta^{n_{i1}} (1 - \theta)^{n_{i2}} \quad (3)$$

121 The true value of  $\theta$  is unknown but can be estimated (as  $\hat{\theta}$ ) by jointly considering all  $m$  samples and  
 122 maximising their summed log-likelihood ( $\mathcal{L}_p$ ):

$$\begin{aligned} \mathcal{L}_p &= \sum_{i=1}^m \left[ \ln \binom{n_{i\bullet}}{n_{i1}} + n_{i1} \ln(\theta) + n_{i2} \ln(1 - \theta) \right] \\ &= \text{constant} + n_{\bullet 1} \ln(\theta) + n_{\bullet 2} \ln(1 - \theta) \end{aligned} \quad (4)$$

123 where  $n_{\bullet 1} \equiv \sum_{i=1}^m n_{i1}$  and  $n_{\bullet 2} \equiv \sum_{i=1}^m n_{i2}$ . Equation 4 is maximised if  $\hat{\theta} = n_{\bullet 1} / (n_{\bullet 1} + n_{\bullet 2})$ . In other words,  
 124 the maximum likelihood solution is found by *pooling* all the counts together. A  $100(1 - \alpha)\%$  confidence interval  
 125 for  $\theta$  is given by:

$$\frac{n_{\bullet 1}}{n_{\bullet 1} + (n_{\bullet 2} + 1)F_{2(n_{\bullet 2} + 1), 2n_{\bullet 1}}^{\alpha}} < \theta < \frac{(n_{\bullet 1} + 1)F_{2(n_{\bullet 1} + 1), 2n_{\bullet 2}}^{\alpha}}{n_{\bullet 2} + (n_{\bullet 1} + 1)F_{2(n_{\bullet 1} + 1), 2n_{\bullet 2}}^{\alpha}} \quad (5)$$

126 where  $F_{a,b}^{\alpha}$  denotes the  $100(1 - \alpha)\%$ -percentile of an F-distribution with  $a$  and  $b$  degrees of freedom. The same  
 127 equation can be used to construct a confidence interval for  $\theta_i$  by replacing  $n_{\bullet 1}$  with  $n_{i1}$  and  $n_{\bullet 2}$  with  $n_{i2}$ .

128

129 Let us consider the binary subcomposition comprised of  $a$  and  $b$  in samples 11 and 13 of Data 1 as an  
 130 example. For sample 11,  $a = b = 34$  so that  $\hat{\theta} = 0.5 \pm 0.11$ . For sample 13,  $a = b = 9$  so that  $\hat{\theta} = 0.5 \pm 0.21$ . In  
 131 other words, although the binomial parameter estimates of these two samples are the same, the width of their  
 132 95% confidence intervals differ by a factor of two. In statistical terms, this is called *heteroscedasticity*. The  
 133 following section of this paper will introduce a graphical means of visualising such data.

### 134 3 A radial plot for point-counting ratios

135 In many Earth Sciences applications, it is not so much the absolute proportions but the ratios between bi-  
 136 nary (sub)compositions that are of interest. For example, in the context of heavy mineral analysis the ap-  
 137 atite/tourmaline, monazite/zircon, and  $\text{TiO}_2$ /zircon ratios (e.g., Morton and Hallsworth, 1994), and the epi-

138 dot/garnet ratio (Heroy et al., 2003) have all be used to indentify the provenance of sediments. In micropalaeon-  
 139 tology, the ratio of benthic to planktonic foraminifera has been used as a productivity index (e.g., Berger and  
 140 Diester-Haass, 1988). And in palynology, the arboreal/non-arboreal pollen ratio is widely used as an index of  
 141 landscape openness (e.g., Herzs Schuh, 2007).

142

143 Given an estimate of the binomial parameter,  $\hat{\theta}$ , such ratios can be simply obtained as  $\hat{\theta}/(1 - \hat{\theta})$ . The cor-  
 144 responding  $100(1 - \alpha)\%$  confidence intervals can be calculated using Equation 5. Applying this to samples 11  
 145 and 13 of Data 1 as an example, we obtain  $a/b$ -ratio estimates of  $1.00 + 0.53/-0.35$  for for sample 11, and  $1.00$   
 146  $+ 1.43/-0.58$  for sample 13. So like the binomial parameter estimates, also the ratio estimates are heteroscedas-  
 147 tic. A radial plot is a graphical device that was specifically designed to display such data.

148

149 Given a set of paired counts  $\{n_{i1}, n_{i2}\}$  (for  $1 \leq i \leq m$ ), the radial plot is a scatter diagram that sets out  
 150  $(z_i - z_o)/s_i$  against  $1/s_i$ , where

$$\begin{aligned} z_i &= \arcsin \left( \sqrt{(n_{i1} + 3/8)/(n_{i\bullet} + 3/4)} \right), \\ z_o &= \arctan \left( \sqrt{n_{\bullet 1}/n_{\bullet 2}} \right), \text{ and} \\ s_i &= 1 / \left( 2\sqrt{n_{i\bullet} + 1/2} \right) \end{aligned} \tag{6}$$

151 Precise measurements plot towards the right hand side of this plot and imprecise measurements to the left.  
 152 The actual  $n_{1i}/n_{2i}$ -ratio is proportional to the slope of a line connecting the  $i^{\text{th}}$  data point to the origin. The  
 153 corresponding values are shown on a circular scale shown at some convenient radial distance away from the origin.

154

155 Figure 1.i shows a radial plot for the first two components ( $a$  and  $b$ ) of Data 1. Samples 11 and 13 of  
 156 Data 1 have been highlighted. Recall that the estimated  $a/b$ -ratio of sample 11 has the same value but twice  
 157 the precision of sample 13. Therefore, sample 11 plots at the same angle, but towards the right of sample 13.

158

159 Figure 1.ii displays the  $a/b$ -ratios of Data 2. Let us have a closer look at samples 2 and 7, which are marked  
 160 in black on this diagram. The total number of grains counted in these two samples are  $n_{2\bullet} = 60 + 24 = 84$ ,  
 161 and  $n_{7\bullet} = 1 + 6 = 7$ , respectively (Table 1). So sample 2 is twelve times larger than sample 7, allowing a more  
 162 precise  $a/b$ -ratio estimate. Sample 2 therefore plots to the right of sample 7, causing the 95%-confidence interval

163 for the former sample to be narrower than that of the latter. Projecting the two samples onto the radial scale  
 164 yields a ratio of  $a/b = 2.5 + 1.7/ - 1.0$  for sample 2 and  $a/b = 0.17 + 1.20/ - 0.16$  for sample 7.

165

166 The dispersion of the data with respect to the counting uncertainties can be visually assessed by comparing  
 167 the vertical scatter of the data with a confidence band of two standard errors wide drawn on either side of the  
 168 origin. If approximately 95% of the data plot within this interval, then the data are compatible with a homo-  
 169 geneous composition. This is the case for Data 1, which can therefore be safely averaged using the pooled ratio  
 170 (Figure 1.i). In contrast, the  $a/b$ -ratios of Data 2 significantly scatter beyond the ‘2-sigma’ confidence region  
 171 (Figure 1.ii). In this case the pooled average should be abandoned in favour of a heterogeneous model (Section 5).

172

173 In conclusion, the radial plot is a useful device to visually assess the dispersion of point-counting ratios. The  
 174 next section of this paper introduces a formal statistical test to make this assessment more quantitative.

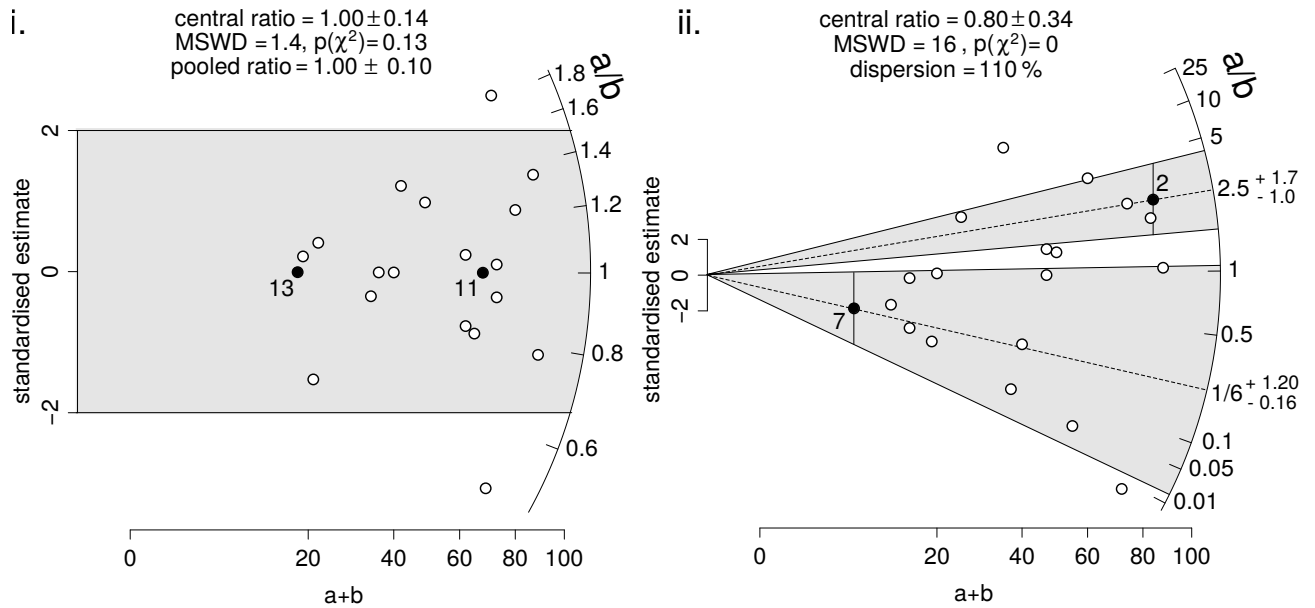


Figure 1: Radial plots of the binary subcompositions  $a$  and  $b$  for Table 1. i) approximately 95% of the samples in Data 1 plot within a symmetric ‘2-sigma’ band around the origin. These data are therefore compatible with a single homogeneous population. The pooled composition then is the best estimate for the average composition (Section 2). ii) the  $a/b$ -ratio of each sample can be obtained by projecting the corresponding scatter point onto the radial scale. The black dots mark samples 2 (right) and 7 (left) of Data 2. Projecting a ‘2-sigma’ error bar onto the same scale yields the 95% confidence intervals of  $a/b = 2.5 + 1.7/ - 1.0$  for sample 2 and  $a/b = 0.17 + 1.20/ - 0.16$  for sample 7, respectively. In contrast with Data 1, the  $a/b$ -ratios of Data 2 do not fit within a ‘2-sigma’ band. In this case the data are more adequately described by a random effects model with two parameters: the central ratio and the dispersion.



## 175 4 A chi-square test for compositional homogeneity

176 In order for the pooled composition to be a meaningful description of the detrital population, all samples must  
177 be derived from a single true composition. In other words, any observed differences in the  $n_{i1}/n_{i2}$ -ratios must  
178 be due to binomial counting statistics alone. The validity of this assumption can be verified by calculating the  
179 chi-square statistic:

$$\chi_{stat}^2 = \frac{1}{n_{\bullet 1} n_{\bullet 2}} \sum_{i=1}^m \frac{[n_{i1} n_{\bullet 2} - n_{i2} n_{\bullet 1}]^2}{n_{i\bullet}} \quad (7)$$

180 If  $\chi_{stat}^2$  is greater than the  $100(1 - \alpha)$ -percentile of a chi-square distribution with  $(m - 1)$  degrees of freedom,  
181 then the null hypothesis of compositional homogeneity is rejected on a  $100(1 - \alpha)\%$  confidence level, where  $\alpha$   
182 is usually taken to be 0.05.

183

184 An alternative way to quantify the dispersion of the data with respect to the expected counting fluctuations  
185 is to divide  $\chi_{stat}^2$  by the number of degrees of freedom. This parameter is known as the ‘reduced chi-square  
186 statistic’, but is also frequently referred to as the ‘Mean Square of the Weighted Deviates’ in the Earth Sciences  
187 (Wendt and Carl, 1991):

$$\text{MSWD} \equiv \chi_{stat}^2 / (m - 1) \quad (8)$$

188 If the observed scatter is entirely due to binomial counting statistics, then the MSWD is expected to take  
189 values close to unity. This is the case for Data 1, which is characterised by an MSWD of 1.4 and a p-value  
190 of 0.13. The latter value is well above the 0.05 cutoff, making the pooled composition the most appropriate  
191 average. The pooled  $a/b$ -ratio of Data 1 is  $1.00 \pm 0.10$ , which agrees with the known ratio of 1.00 (Section 1).

192

193 MSWD values significantly greater than one indicate the presence of excess scatter beyond the binomial  
194 counting statistics. This is the case for Data 2, which yields an MSWD-value of 16 and a p-value close to zero.  
195 In this situation, the pooled average is not the most appropriate estimator for the population average and a  
196 more realistic model must be used. An example of one such model is given in Section 5.

## 5 Continuous mixtures

Datasets that fail the chi-square test for homogeneity are incompatible with a single binomial population. Instead their binomial population parameter  $\theta$  may be drawn from a continuous distribution. Suppose that  $\theta$  is drawn from a logistic normal distribution with geometric mean  $\mu$  and coefficient of variation  $\sigma$ , and define

$$\beta \equiv \ln \left( \frac{\theta}{1-\theta} \right) \sim \mathcal{N}(\mu, \sigma^2) \quad (9)$$

where  $\mathcal{N}(\mu, \sigma^2)$  stands for “the normal distribution with mean  $\mu$  and variance  $\sigma^2$ ”. Note that  $\beta$  is a logratio similar to those defined in Equation 1. Given the usual  $m$  sets of point-counts  $\{n_{i1}, n_{i2}\}$  and maximising the log-likelihood function  $\mathcal{L}_c$

$$\mathcal{L}_c = \sum_{i=1}^m \ln \left\{ \binom{n_{i\bullet}}{n_{i1}} \int_{-\infty}^{\infty} \frac{\exp[\beta n_{i1}]}{(\exp[\beta] + 1)^{n_{i\bullet}}} \frac{\exp \left[ -\frac{1}{2} \left( \frac{\beta - \mu}{\sigma} \right)^2 \right]}{\sigma \sqrt{2\pi}} d\beta \right\} \quad (10)$$

yields two estimates  $\hat{\mu}$  and  $\hat{\sigma}$  whose approximate standard errors may be obtained by inverting the Hessian matrix of second derivatives of  $\mathcal{L}_c$ . The integrals in Equation 10 cannot be evaluated analytically, but a quick numerical solution is provided by Galbraith and Laslett (1993). The ‘central’ composition is then defined as:

$$\hat{\theta} = \frac{\exp[\hat{\mu}]}{\exp[\hat{\mu}] + 1} \quad (11)$$

which is akin to the inverse logratio transformation defined in Equation 2. The central  $a/b$ -ratio for Data 2 is  $0.80 \pm 0.34$  (Figure 1.ii), which again agrees with the true ratio of 1.00 that was reported in Section 1.

The dispersion estimate  $\hat{\sigma}$  quantifies the geological variability of the underlying population. This is just as useful a quantity as the central value itself. It estimates the relative spread of the underlying population without the binomial counting errors. For example the coefficient of variation (standard deviation divided by mean) of the  $a/b$ -ratio measurements for Data 2 is  $\sim 260\%$ . This is far greater than the  $\sim 110\%$  dispersion estimated by the random effects model. The latter estimate is much closer to the true dispersion of the underlying population, whose value is 100% (Section 1).

It is useful to note that, for samples that pass the chi-square test for sample homogeneity, the pooled ratio is the same as the central ratio. This is indeed the case for Data 1, which yields a pooled ratio of  $1.00 \pm 0.10$

219 and a central ratio of  $1.00 \pm 0.14$ . The larger 95% uncertainty interval of the latter is due to the loss of one  
 220 degree of freedom that is required to estimate  $\hat{\sigma}$ .

## 221 6 Ternary models

222 The statistical models presented in the previous sections can be generalised from two to three or more com-  
 223 ponents. This is trivial for homogeneous populations such as Data 1, whose pooled composition is shown in  
 224 Figure 2.i. For heterogeneous populations such as Data 2, the continuous mixture model of Section 5 can be  
 225 generalised by defining two population parameters  $\beta_1 \equiv \ln[\theta_1] - \ln[1 - \theta_1 - \theta_2]$  and  $\beta_2 \equiv \ln[\theta_2] - \ln[1 - \theta_1 - \theta_2]$ .  
 226 Assuming that  $\beta_1$  and  $\beta_2$  are drawn from a bivariate normal distribution with mean  $M$  and covariance matrix  
 227  $\Sigma$ , the three-component equivalent to Equation 10 becomes

$$\mathcal{L}_t = \sum_{i=1}^m \ln \left\{ \frac{n_{i\bullet}!}{n_{i1}!n_{i2}!n_{i3}!} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\exp[\beta_1 n_{i1} + \beta_2 n_{i2}]}{(\exp[\beta_1] + \exp[\beta_2] + 1)^{n_{i\bullet}}} \right. \\ \left. \frac{\exp\left[-\frac{1}{2}(B-M)^T \Sigma^{-1}(B-M)\right]}{2\pi\sqrt{|\Sigma|}} d\beta_1 d\beta_2 \right\} \quad (12)$$

228 where  $\{n_{i1}, n_{i2}, n_{i3}\}$  are the ternary counts of the  $i^{\text{th}}$  sample, with  $n_{i1} + n_{i2} + n_{i3} = n_{i\bullet}$  for  $1 \leq i \leq m$ ;

$$B = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}; M = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \text{ and } \Sigma \equiv \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}$$

229 in which  $\sigma_1$  and  $\sigma_2$  are the standard deviations of  $\beta_1$  and  $\beta_2$ , and  $\sigma_{1,2}$  is their covariance. Equation 12, like  
 230 Equation 10, does not have an analytical solution and requires numerical integration for each sample. The fast  
 231 algorithm of Galbraith and Laslett (1993) can be used to estimate  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$  and  $\sigma_2$  so that only the covariance  
 232  $\sigma_{1,2}$  remains to be found. The central composition is then estimated by substituting  $\hat{\mu}_1$  for  $u_i$  and  $\hat{\mu}_2$  for  $v_i$  in  
 233 Equation 2. Figure 2.ii applies this model to Data 2, showing the central composition as a black square and  
 234 using the dispersion estimate  $\hat{\Sigma}$  to define a 95% confidence region for the underlying population (red line).

235

236 The definition of  $\beta_1$  and  $\beta_2$  that is used in Equation 12 is consistent with the logratio approach of Equation 1.  
 237 However other parameterisations are possible as well. For example, we could define three logistic population  
 238 parameters ( $\beta_i \equiv \ln[\theta_i] - \ln[1 - \theta_i]$  for  $1 \leq i \leq 3$ ) to ensure compatibility with the bivariate random effects

239 model of Section 5. These alternative parameterisations are interchangeable with each other and can easily be  
 240 converted to each other.

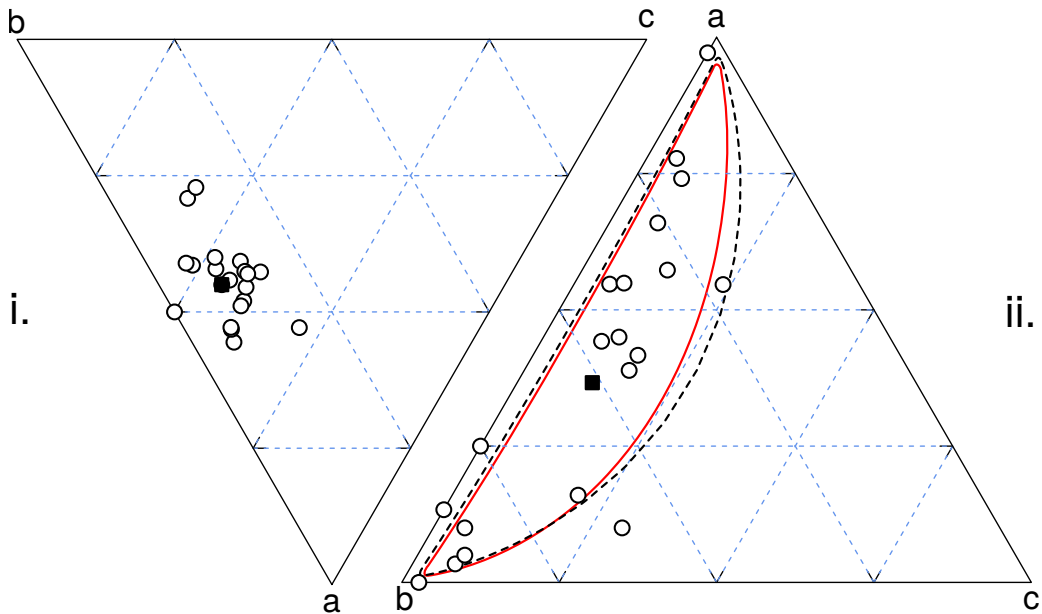


Figure 2: Statistical analysis of ternary point-counting data. i. Data 1 is sampled from a homogeneous population, whose mean is given by the pooled composition shown as a black square. ii. Data 2 is derived from a continuous mixture (see the caption of Table 1 for details). Its estimated central composition is shown as a black square, and the 95% confidence envelope corresponding to its dispersion parameters as a red contour line. The dashed black contour marks the true dispersion region for the population, also shown at 95% confidence.

## 241 7 Correspondence analysis

242 The previous sections of this paper have shown that binary or ternary datasets can be visualised as radial plots  
 243 and ternary diagrams, respectively. These two-dimensional graphics are useful for interpreting point-counting  
 244 data, but cannot so easily be applied to higher dimensional datasets. In this section, we will consider the general  
 245 case of a  $K$ -component dataset  $X$  contained in an  $[m \times K]$ -matrix. We will explore some strategies to display  
 246 such a dataset as a two-dimensional graphic.

247

248 Principal Component Analysis (PCA, Pearson, 1901) is an ordination techniques that is commonly used for  
 249 exploratory data analysis of multi-dimensional datasets. PCA is a two step process. First, the data are ‘centred’  
 250 by subtracting the arithmetic mean composition from each column. Second, the centred data are decomposed  
 251 into an orthogonal set of  $K$  principal components. Plotting the first two principal components against each

252 other then yields the desired two-dimensional data projection.

253

254 Unfortunately, PCA cannot readily be applied to compositional data or point-counting data. This is because  
255 the first step involves taking an arithmetic mean, which we have already shown to be problematic in Section 1.  
256 Subjecting the data to a logratio transformation prior to PCA analysis solves this problem (Aitchison, 1983).  
257 But this solution generally does not work for point-counting data due to its inability to handle zero count  
258 data. This issue is aggravated by the tendency for high dimensional datasets to contain more zeros than lower  
259 dimensional datasets do.

260

261 Correspondence Analysis (CA, Greenacre, 1984) fixes these issues by explicitly treating the data as counts.  
262 CA is a multivariate ordination technique that is conceptually similar to PCA. To understand the relationship  
263 between the two methods, it is useful to point out that PCA and CA are both special case of another exploratory  
264 data analysis method called Multidimensional Scaling (MDS, Kruskal and Wish, 1978; Vermeesch, 2013). Given  
265 a table of pairwise ‘dissimilarities’ between samples, MDS produces a map in which similar samples plot close  
266 together and dissimilar samples plot far apart.

267

268 PCA is a special case of MDS in which the dissimilarities are Euclidean distances. CA is another special  
269 case of MDS in which the dissimilarities are chi-square distances (Legendre and Gallagher, 2001; Greenacre,  
270 2005):

$$d_{ij} = \sqrt{\sum_{k=1}^K \frac{X_{\bullet\bullet}}{X_{\bullet k}} \left( \frac{X_{ik}}{X_{i\bullet}} - \frac{X_{jk}}{X_{j\bullet}} \right)^2} \quad (13)$$

271 where  $d_{ij}$  is the dissimilarity between samples  $i$  and  $j$  (with  $1 \leq i, j \leq m$ );  $X_{\bullet k} = \sum_{i=1}^m X_{ik}$ ;  $X_{i\bullet} = \sum_{k=1}^K X_{ik}$ ;  
272  $X_{j\bullet} = \sum_{k=1}^K X_{jk}$ ; and  $X_{\bullet\bullet} = \sum_{i=1}^m \sum_{k=1}^K X_{ik}$ . In the case of PCA, the principal components are obtained by  
273 linear combination of the original variables. The weightings of these variables can be displayed together with  
274 the transformed data as a biplot (Aitchison and Greenacre, 2002). The same principle can be applied to CA  
275 (Figure 3).

## 276 8 Examples

277 All the methods discussed in this paper were added to the `provenance` package of Vermeesch et al. (2016).  
278 Written in the statistical programming language R, `provenance` comes with a query-based user interface that  
279 does not require any programming skills. Alternatively, the full functionality of the package can also be accessed  
280 via the command line, as demonstrated in the following tutorial.

281

282 Point-counting data can be read from a `.csv` file using the `read.counts` function. For example, to read the  
283 second dataset from Table 1:

```
284 data2 <- read.counts('data2.csv')
```

285 Plotting the ratios of the first two variables ( $a/b$ ) as a radial plot (Figure 1.i):

```
286 radialplot(data2,num='a',den='b')
```

287 where `num` and `den` are optional arguments denoting the names of the numerator and denominator compo-  
288 nent, respectively. Plotting the full dataset on a ternary diagram and constructing its 95% confidence region  
289 (Figure 2.ii):

```
290 # create a ternary data object:
```

```
291 tern <- ternary(data2)
```

```
292 # show the data on a ternary diagram
```

```
293 # as white circles without data labels:
```

```
294 plot(tern,pch=1,labels=NA)
```

```
295 # add the 95% confidence region:
```

```
296 ternary.ellipse(tern,alpha=0.05)
```

297 where everything that follows a hash character (`#`) is a comment and is ignored. Next, let us consider a real  
298 dataset of heavy mineral counts from Namibia published by Vermeesch et al. (2016). The following code snippet  
299 calculates the central composition for these data:

```
300 HM <- read.counts('HM.csv')
```

```
301 avg <- central(HM)
```

302 The variable `avg` contains a  $[5 \times 15]$  table in which each column corresponds to a mineral species, and the  
303 rows contain (1) the central value for the binomial parameters ( $\theta_i$  for  $1 \leq i \leq 15$ ) for those minerals; (2) the

304 standard error for the binomial parameters; (3) the overdispersion parameter for the binary composition pa-  
305 rameter  $\ln[\theta_i] - \ln[1 - \theta_i]$ ; (4) the MSWD value for each binary subcomposition; and (5) the corresponding p-value.

306

307 Next, we will perform a correspondence analysis of the Namib data. But before doing so, it is important  
308 to point out that CA is most sensitive to the least abundant components. To mitigate the effects of this  
309 phenomenon, it is useful to pre-process the data. The following code snippet selects the most abundant minerals  
310 (epidote, garnet, amphibole and clinopyroxene) from the datasets and amalgamates the ultra-stable minerals  
311 (zircon, tourmaline and rutile), which have similar petrological significance:

```
312 HM2 <- amalgamate(HM,ztr=c('zr','tm','rt'),ep='ep',  
313                       gt='gt',amp='amp',cpx='cpx')
```

314 The resulting data object (HM2) still contains a number of zero values, but is no longer dominated by them.  
315 The actual CA calculation then proceeds as follows:

```
316 # perform the calculations:  
317 ca <- CA(HM2)  
318 # show the results as a biplot:  
319 plot(ca)
```

320 The biplot (Figure 3) displays the samples in black and the minerals as red arrows. The tight clustering  
321 of samples N1, N2, N3, N10, N12, N14, T8 and T13 reflects the compositional similarity between these sam-  
322 ples, which were all derived from the coastal parts of the Namib Sand Sea (Vermeesch and Garzanti, 2015).  
323 In contrast, inland samples N4, N5, N8 and N9 plot elsewhere, indicating that they have a different compo-  
324 sition. This is due to a combination of provenance and hydraulic sorting effects (Vermeesch and Garzanti, 2015).

325

326 The configuration of the mineral labels provides further insight into the factors that cause the dispersion  
327 of the samples on the biplot. For example, the orientation of the red arrows shows that samples N8 and  
328 N9 are enriched in garnet and ultra-stable minerals, whereas sample N5 is enriched in epidote relative to the  
329 coastal samples. The arrows for epidote and clinopyroxene point in opposite directions, indicating that these  
330 two minerals are anti-correlated with each other. In contrast, the arrow for garnet is perpendicular to that of  
331 epidote. This indicates that garnet and epidote are uncorrelated with each other.

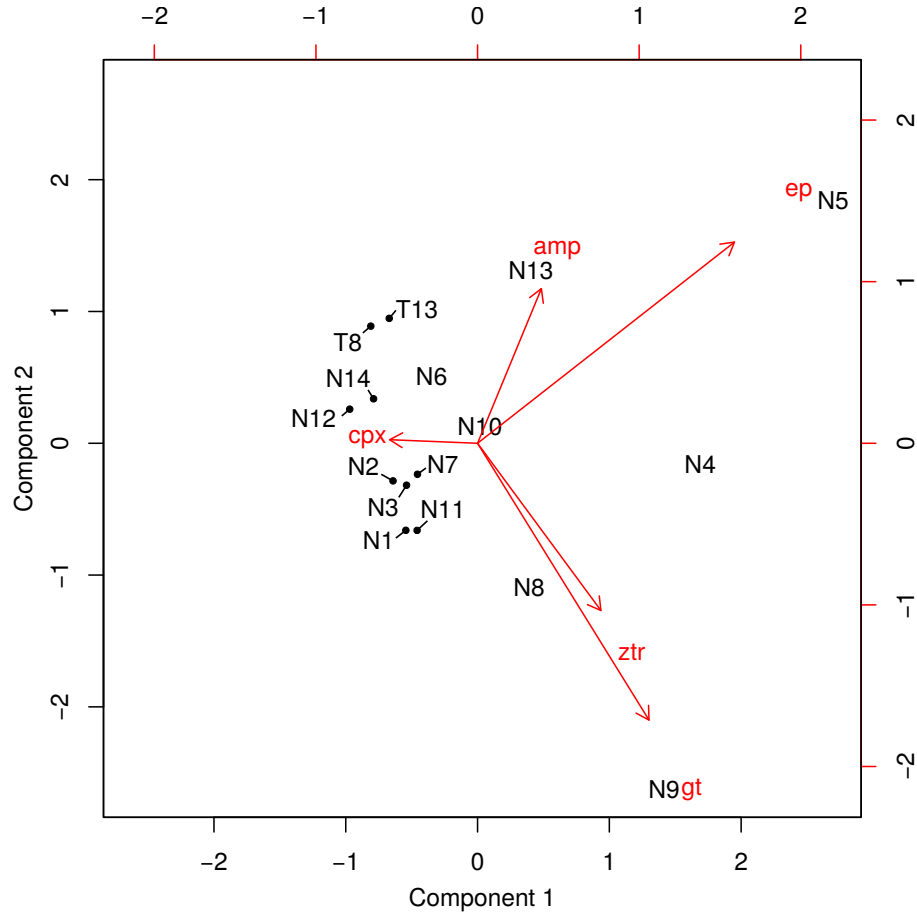


Figure 3: Correspondence analysis of heavy mineral compositions from Namibia (Vermeesch et al., 2016) shown as a biplot. Samples are shown in black, minerals in red.

## 332 9 Discussion and conclusions

333 It is common practice in sedimentary petrography, palaeontology and palynology to report the relative abun-  
 334 dances of minerals, fossils or pollen as percentages. Unfortunately, by doing so one loses the ability to quantify  
 335 the statistical uncertainty of the underlying point-counting data. Normalisation of point-counts also compro-  
 336 mises the ability to deal with missing (zero) components.

337

338 The statistical methods reviewed in this paper are built on the recognition that point-counts represent a  
 339 distinct class of data. This new data class shares aspects with, but is fundamentally different from Aitchison  
 340 (1986)'s compositional data. Compositional data only carry relative information (Pawlowsky-Glahn and Buc-  
 341 cianti, 2011), and the absolute abundances of their components are irrelevant. In stark contrast with this, for  
 342 point-counting data the absolute abundances *do* matter, because they control the precision of the estimated



343 compositions (Bloemsma and Weltje, 2015).

344

345 This observation leads to a first recommendation, which is to report the total number of counts for each  
346 sample in published data tables. This allows the recovery of the raw point-counting data. Those data can then  
347 be further analysed using the techniques introduced in this paper.

348

349 Compositional data and point-counting data are closely related to each other. In fact, point-counting *data*  
350 are underlain by compositional *populations*. These populations can be constrained using a combination of multi-  
351 nomial and logratio statistics.

352

353 If the data are underlain by a single, fixed composition, then the point-counting data follow a multinomial  
354 distribution. In this case, the fixed composition of the underlying population can be estimated by pooling all  
355 the data together (Section 2). However, this simple scenario rarely occurs in the real world. Provided that  
356 a dataset is large enough, virtually all populations are overdispersed with respect to the multinomial point-  
357 counting uncertainties (Section 4).

358

359 For two-component systems, the degree of overdispersion can be visually assessed on a radial plot (Section 3).  
360 The dispersion may then be quantified using a three parameter continuous mixture model (Section 5), which  
361 can be generalised to three (or more) components (Section 6). The continuous mixture model assumes that  
362 the point-counting data are underlain by a logistic normal distribution. Although more realistic than the ho-  
363 mogeneous population assumed by the pooled composition, the continuous mixture model is still a very simple  
364 approximation to real geological scenarios.

365

366 Correspondence Analysis was introduced as an effective tool for exploratory analysis of more complex and  
367 higher dimensional datasets (Section 7). It does not seek to capture the data in a simplified analytical form.  
368 Instead, CA distills the salient similarities and differences between samples as a two-dimensional ‘map’, and  
369 which the variables can also be shown. Such biplots can provide valuable geological insights that would be  
370 difficult to obtain otherwise.

371

372 CA is closely related to Principal Component Analysis. PCA can be applied to compositional data and  
373 uses Aitchison’s Euclidean logratio-distance as a measure to compare the (dis)similarities between samples. In

374 contrast, CA uses the chi-square distance (Equation 13), which makes it immune to the zero-count problem.  
375 Once we recognise the close affinity between the Aitchison distance and compositional data on the one hand, and  
376 between the chi-square distance and point-counting data on the other hand, then it is possible to add further  
377 complexity to our statistical analysis.

378

379 For example, Vermeesch and Garzanti (2015) introduce a technique called 3-way multidimensional scaling  
380 to combine different datasets together for the purpose of sedimentary provenance analysis. Using the insights  
381 gained from this paper, we could use the Aitchison distance to compare the major and trace element compositions  
382 of different samples, and the chi-square distance to compare their bulk petrography and heavy mineral counts.

## 383 Acknowledgments

384 The methods presented in this paper follow from the work of Rex Galbraith, who is gratefully acknowledged  
385 for patiently answering numerous questions from the author. Gert-Jan Weltje and an anonymous reviewer are  
386 thanked for their constructive comments on the submitted manuscript.

## 387 References

- 388 Aitchison, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society*, 44:139–177,  
389 1982.
- 390 Aitchison, J. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983. doi:  
391 10.1093/biomet/70.1.57.
- 392 Aitchison, J. *The statistical analysis of compositional data*. London, Chapman and Hall, 1986.
- 393 Aitchison, J. and Greenacre, M. Biplots of compositional data. *Journal of the Royal Statistical Society: Series*  
394 *C (Applied Statistics)*, 51(4):375–392, 2002.
- 395 Barkley, F. A. The statistical theory of pollen analysis. *Ecology*, 15(3):283–289, 1934.
- 396 Berger, W. and Diester-Haass, L. Paleoproductivity: the benthic/planktonic ratio in foraminifera as a produc-  
397 tivity index. *Marine Geology*, 81(1-4):15–25, 1988.

- 398 Bloemsma, M. R. and Weltje, G. J. Reduced-rank approximations to spectroscopic and compositional data:  
399 A universal framework based on log-ratios and counting statistics. *Chemometrics and Intelligent Laboratory*  
400 *Systems*, 142:206–218, 2015.
- 401 Buzas, M. A. Another look at confidence limits for species proportions. *Journal of Paleontology*, 64(5):842–843,  
402 1990.
- 403 Clark, R. L. Point count estimation of charcoal in pollen preparations and thin sections of sediments. *Pollen et*  
404 *spores*, 1982.
- 405 Dryden, A. Accuracy in percentage representation of heavy mineral frequencies. *Proceedings of the National*  
406 *Academy of Sciences*, 17(5):233–238, 1931.
- 407 Fatela, F. and Taborda, R. Confidence limits of species proportions in microfossil assemblages. *Marine Mi-*  
408 *cropaleontology*, 45(2):169–174, 2002.
- 409 Galbraith, R. F. *Statistics for fission track analysis*. CRC Press, 2005.
- 410 Galbraith, R. Graphical display of estimates having differing standard errors. *Technometrics*, 30(3):271–281,  
411 1988.
- 412 Galbraith, R. and Laslett, G. Statistical models for mixed fission track ages. *Nuclear tracks and radiation*  
413 *measurements*, 21(4):459–470, 1993.
- 414 Greenacre, M. Weighted metric multidimensional scaling. In *New developments in classification and data*  
415 *analysis*, pages 141–149. Springer, 2005.
- 416 Greenacre, M. J. Theory and applications of correspondence analysis. page 364, 1984.
- 417 Heroy, D. C., Kuehl, S. A., and Goodbred Jr, S. L. Mineralogy of the Ganges and Brahmaputra Rivers:  
418 implications for river switching and Late Quaternary climate change. *Sedimentary Geology*, 155(3-4):343–  
419 359, 2003.
- 420 Herzschuh, U. Reliability of pollen ratios for environmental reconstructions on the Tibetan Plateau. *Journal of*  
421 *Biogeography*, 34(7):1265–1273, 2007.
- 422 Kruskal, J. B. and Wish, M. *Multidimensional scaling*, volume 07-011 of *Sage University Paper series on*  
423 *Quantitative Application in the Social Sciences*. Sage Publications, Beverly Hills and London, 1978.

- 424 Legendre, P. and Gallagher, E. D. Ecologically meaningful transformations for ordination of species data.  
425 *Oecologia*, 129(2):271–280, 2001.
- 426 Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. Dealing with zeros and missing values  
427 in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.
- 428 Morton, A. C. and Hallsworth, C. Identifying provenance-specific features of detrital heavy mineral assemblages  
429 in sandstones. *Sedimentary Geology*, 90(3-4):241–256, 1994.
- 430 Patterson, R. T. and Fishbein, E. Re-examination of the statistical methods used to determine the number of  
431 point counts needed for micropaleontological quantitative research. *Journal of Paleontology*, 63(2):245–248,  
432 1989.
- 433 Pawlowsky-Glahn, V. and Buccianti, A. *Compositional data analysis: Theory and applications*. John Wiley &  
434 Sons, 2011.
- 435 Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin*  
436 *Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- 437 Van der Plas, L. and Tobi, A. A chart for judging the reliability of point counting results. *American Journal of*  
438 *Science*, 263(1):87–90, 1965.
- 439 Vermeesch, P. Multi-sample comparison of detrital age distributions. *Chemical Geology*, 341:140–146, 2013.
- 440 Vermeesch, P. and Garzanti, E. Making geological sense of ‘Big Data’ in sedimentary provenance analysis.  
441 *Chemical Geology*, 409:20–27, 2015.
- 442 Vermeesch, P., Resentini, A., and Garzanti, E. An R package for statistical provenance analysis. *Sedimentary*  
443 *Geology*, 2016.
- 444 Weltje, G. Quantitative analysis of detrital modes: statistically rigorous confidence regions in ternary diagrams  
445 and their use in sedimentary petrology. *Earth-Science Reviews*, 57(3-4):211 – 253, 2002. ISSN 0012-8252. doi:  
446 DOI: 10.1016/S0012-8252(01)00076-9.
- 447 Wendt, I. and Carl, C. The statistical distribution of the mean squared weighted deviation. *Chemical Geology:*  
448 *Isotope Geoscience Section*, 86(4):275–285, 1991.

449 Weng, C., Hooghiemstra, H., and Duivenvoorden, J. F. Challenges in estimating past plant diversity from  
450 fossil pollen data: statistical assessment, problems, and possible solutions. *Diversity and distributions*, 12(3):  
451 310–318, 2006.