Earth and Planetary Science Letters 501 (2018) 112–118

Contents lists available at ScienceDirect

Earth and Planetary Science Letters

www.elsevier.com/locate/epsl

Statistical models for point-counting data

Pieter Vermeesch

Department of Earth Sciences, University College London, United Kingdom

ARTICLE INFO

Article history:
Received 27 April 2018
Received in revised form 3 August 2018
Accepted 8 August 2018
Available online xxxx
Editor: A. Yin

Keywords:
statistics
point-counting
heavy mineral analysis
petrography
micropalaeontology
palynology

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 magnitude of these two sources of dispersion can be assessed by a chi-square statistic and 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 silicilastic 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.

E-mail address: p.vermeesch@ucl.ac.uk.

https://doi.org/10.1016/j.epsl.2018.08.019

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:

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

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

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

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

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

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

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

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

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.

Data 1					Data 2				
#	*a*	*b*	*c*	*R*	#	*a*	*b*	*c*	*R*
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
C	541	539	142	$N = 1222$	*C*	433	452	90	$N = 987$

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

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

All the techniques discussed above will be illustrated with a combination of synthetic and real examples. The methods of Sections 2–6 will use the two datasets shown in Table 1. Data 1 consists of 20 random samples of 23–94 items each, which were drawn from a discrete trinomial distribution with 45% of component *a*, 45% of component *b* and 10% of component *c*. Data 2 comprises a further 20 samples that were drawn from a continuous distribution whose mode is the same as that of Data 1, but which adds 100% of dispersion around this mode. Thus, Data 2 has two sources of dispersion (counting error and true population dispersion), whereas Data 1 only has one (counting error). Note that both datasets contain fewer counts per sample than is customary in real world applications. But they are nevertheless realistic if we consider them to be ternary subcompositions of higher dimensional datasets.