Tools for identifying unexpectely low microfossil count sums

Richard J. Telford*,[a]

[a] *Department of Biological Sciences, University of Bergen and Bjerknes Centre for Climate Research, Post Box 7803, N-5020 Bergen, Norway*

Abstract

Microfossil counts are a key data type in palaeoecology. Recent work has raised the possibility that some authors might misreport an important quality control parameter, the counts sums, occasionally dramatically so. This paper introduces methods that can flag assemblages with potentially misreported count sums and finds that some assemblage datasets that fail these tests.

Keywords: Microfossil counts; quantitative methods

1. Introduction

Six percent of papers published in *Molecular and Cellular Biology* show evidence of inappropriate image duplication (Bik et al. 2018) either due to error or, more rarely, misconduct. It would be reckless to assume that the palaeoecological literature does not have equivalent problems: the low rate of retractions in the ecological and geological literature (Grieneisen and Zhang 2012) may partially reflect our limited ability to detect errors and misconduct rather than their prevalence.

Several numerical tools have been developed to identify questionable data. Deviations from the distribution of digits expected from the Newcomb-Benford law has been used to detect issues with scientific and financial data (Barabesi et al. 2018). Carlisle (2017) identified papers where the baseline differences in means for different treatments were surprisingly high or low given the variance of the data. Brown et al. (2017) developed the granularity-related inconsistency means (GRIM) test of whether means of integer data are consistent with the reported sample size. However none of these tests are directly applicable to microfossil assemblage data, one of the most common types of palaeoecological data.

A common assertion in papers reporting microfossil assemblage data is that a minimum of N microfossils were counted, where N is often fifty for chironomids and several hundred for pollen and diatoms. This is important because larger count sums are associated with smaller uncertainties, both in relative abundance of taxa and derived statistics such are transfer function reconstructions (Heiri and Lotter 2001). However, mistakes happen and metadata such as count sums

*Corresponding Author
Email address: richard.telford@uib.no (Richard J. Telford)

can be forgotten once percentages are calculated. In addition, given the time-consuming nature of microfossil counting, especially when preservation is poor or concentrations are low, there may be an incentive to misreport the minimum count sum. The risk that count sums will be mis-reported is not just theoretical: Larocque-Tobler et al (2015) reported that their chironomid count sums were at least fifty head capsules; a subsequent corrigendum (Larocque-Tobler et al. 2016) acknowledged that count sums were actually as low as nineteen. Telford (2019a) reports some other cases where count sums may be much lower than reported.

If the genuine assemblage counts are archived for all taxa, it is trivial to identify undercounts. Unfortunately, count data might be falsified so that it appears to meet the reported count sum. A more common problem is that, regrettably, many palaeoecologists archive percent data without an indication of the count sum. This paper develops simple tests that can flag if the count data might have been misrepresented, or if the count sum of percent data is perhaps lower than reported.

The key insight that allows inference about the count sum is that assemblage data are expected to follow the typical features of a rank abundance curve. In particular, because there are many rare taxa in most communities (Darwin 1859), most community or assemblage samples will include taxa represented by a single individual, hereafter singletons, unless the sampling effort is high relative to the taxonomic richness (Coddington et al. 2009). In the occasional assemblages without singletons, the greatest common divisor should usually be one, i.e. few assemblages should have all counts divisable by an integer larger than one, especially if species richness is high.

In general, it is not possible to determine the sum from which percent were calculated, but the properties of community and assemblage counts make it possible to estimate it. Given that we expect the rarest taxon to be a singleton, the count sum N can be estimated as $1/p_{min} \times 100$ where p_{min} is the percent abundance of the rarest taxon. This method will fail for assemblages without singletons. We also expect the percent p to be calculated from integer counts, therefore, there should be a count sum N such that $p_i/100 \times N$ is, within rounding error, an integer for all i taxa. Possible values for N can be found by a direct search algorithm over the range of plausible values of N. An infinite number of possible count sums that are consistent with the percent exist, but the lowest will give the correct value of N except in cases where the greatest common divisor is greater than one. These tests are closely related to GRIM (Brown and Heathers 2017), as all rely on the granularity of percentages calculated from integer data.

This paper aims to test the methods presented above and present some cases with unexpected results. Some complications and caveats are discussed.

2. Methods

Publicly available datasets were downloaded from the Palaeodata Center, Neotoma, Pangaea and other sources. A range of ecological and palaeoecological

Table 1: Percent of bird counts without singletons and of greatest common divisor (GCD) greater than one of counts without singletons, at different taxonomic levels.

Taxonomic level	Percent without singletons	Percent GCD > 1	Mean richness
species	0.05	0.00	53
genus	0.15	0.00	45
family	4.31	0.07	24
order	17.38	3.37	9

data were sought to allow for difference in typical count sum and species richness. Fossil pollen assemblage in the Neotoma database (Williams et al. 2018) (downloaded 2019-05-03) with small count sums (< 50), that appeared to be percentages, or were documented as back-transformed from digitised data were excluded. Datasets with percent data where the minimum count sum was not reported in the associated paper were excluded. Datasets discussed by Telford (2019a) as possibly having under-reported count sums were also excluded to avoid double reporting. Datasets with possible mis-reporting are anonymised, but no attempt is made to diagnose whether errors or misconduct are responsible. This paper does not attempt to be an exhaustive survey of all the data available.

All analyses were done in R version 3.4.4 (R Core Team 2018) and used the packages extraDistr version 1.8.10 (Wolodzko 2018), numbers version 0.7.1 (Borchers 2018), and countSum 0.0.3 (Telford 2019b). Code to replicate all the analyses shown above is archived at https://github.com/richardjtelford/count.check.ms.

3. Results

3.1. Prevalence of singletons

A tiny minority of the over 65,000 bird counts from the North American breeding bird survey (Pardieck et al. 2018) lack singletons at the species level (Table 1)). To explore the effect of taxonomic richness on the prevalence of singletons, I aggregate the birds counts to progressively lower taxonomic resolutions. As richness declines, the proportion of counts lacking singletons increases (Table 1), reaching a moderate proportion at the order level, where most of the counts lacking singletons have fewer than five taxa. The vast majority of counts without singletons have a greatest common divisor of one, except when taxonomic richness is low.

To test the sensitivity of the prevalence of counts without singletons to the count sum, I re-sample the counts with 400 or more observations to smaller count sums using a multivariate hypergeometric distribution. At the order level, the proportion of counts without singletons declines steeply until it reaches a minimum at about 200, thereafter is rises slowly (Fig. 1). The prevalence of counts without singletons is high for small counts because it is possible that only the common taxa are counted; with larger counts, the chance of counting a single individual of a rare taxon increases. With even larger counts the prevalence of counts without singletons increases due to saturation as singletons become doubletons and few new taxa are found.

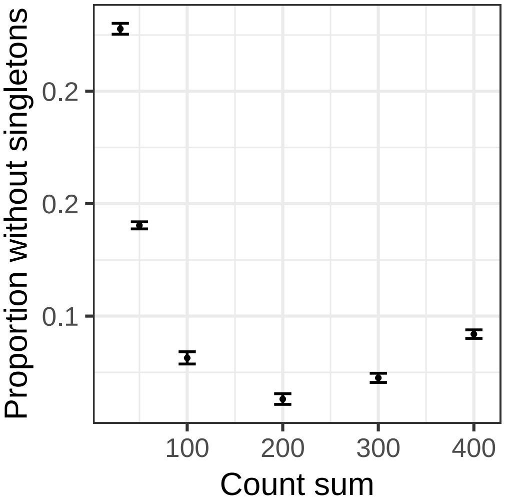

Figure 1: Effect of count sum on the proportion of bird counts without singleton orders. Results are the mean of ten trials, error bars are two standard errors.

The 862 diatoms counts from Owen's Lake span a wide range of count sums (1 – 701) as, if concentrations were low, the requirement that at least 300 valves were counted, was replaced by a minimum area of slide to count (Bradbury 1997). Discounting the 20 mono-specific assemblages (all with very low count sums), 97.4% of assemblages have singletons with a median number of singletons per assemblage of five. Assemblages with count sums above 50 have a higher chance of having singletons (98.2% vs 89%). 99.5% of assemblages have a greatest common divisor of one, the remainder have a greatest common divisor of two. Assemblages with a greatest common divisor above one all have low taxonomic richness.

The 1775 assemblages in the North American testate amoeba training set with available count data (Amesbury et al. 2018) have count sums between 52 and 456 tests (median = 151), and taxonomic richness ranges between 2 and 31 (median = 15) taxa. 94.9% of the assemblages have at least one singleton (median 4) and 99.7% of assemblages have a greatest common divisor of one. Of the five assemblages with a greatest common divisor above one, three have fewer than five taxa; four have a greatest common divisor of two, and the other one has a greatest common divisor of three.

The pollen data from the Neotoma database included over 182 thousand assemblages in 4130 datasets. Nearly all assemblages have singletons (97%) and a greatest common divisor of one (99.6%). The assemblages with a greatest common dividor above one are not randomly distributed among the datasets: 96.6% of the datasets have no such assemblages. Some data sets with a high proportion of assemblages with a greatest common divisor above are older datasets which may, despite the available metadata, have been digitised with the loss of rare taxa and precision. For example, one dataset includes 84 assemblages with a median richness of 16 taxa. Count sums vary between 69 and 3201. The greatest common divisor is three for all assemblages, that is all 1300 counts in the dataset are divisible by three. Excluding such datasets would further increase the proportion of assemblages with singletons. Some other Neotoma datasets are discussed is a subsequent section.

3.2. Estimating the percent sum

The Owen's Lake dataset (Bradbury 1997) also includes percentages for each taxa, allowing the count sums estimated from the percent data to be verified. Excluding monospecific assemblages, for which no meaningful estimate of the count sum is possible, the minimum percent method and the direct search method correctly estimate the count sum for, respectively, 97.4% and 99.5% of the assemblages. The few assemblages that fail the direct seach method are species poor and have low count sums.

The 22 chironomid head capsule assemblage dataset from Last Chance Lake (Axford et al. 2017) also includes both the count sums and the percentage of each taxa. The percentages are given to two decimal places: to account for rounding errors, the countSum package uses the smallest and largest values consistent with the reported percentage. With the minimum percent method the estimated count sums are, within error, either identical to the reported count sum or twice as much (Fig. 2). With the direct search method, the estimated count sums are all exactly twice the reported count sum. The factor of two difference is because all the counts include half head capsules (but only sometimes is the rarest taxon represented by a half head capsule). Reporting half microfossils is common for several microfossil groups, including chironomids and pollen (especially for bisaccate conifers); occasionally other fractions are reported. Half counts make the estimated count too high by a factor of two, so do not risk incorrectly flagging count sums as being too small. The direct search method is more precise than the minimum percent method as the rounding error is relatively smaller on the larger percent values the method uses.

3.3. Unexpected count data

Some of the pollen data in the Neotoma database have an inexplicably high proportions of assemblages with a greatest common divisor greater than one. Here I focus on four datasets produced by the same research group. All the datasets, which are relatively recent, include information on the spike used to calculate pollen concentration, so cannot have been digitised. The four datasets have some assemblages with a greatest common divisor of two interspersed, sometimes regularly, amongst assembages with the expected greatest common divisor of one. The assemblages with a greatest common divisor of one typically have several singletons (Table 2)).

Assemblages with a greatest common divisor of one typically have a higher taxonomic richness than assemblages with a greatest common divisor of two

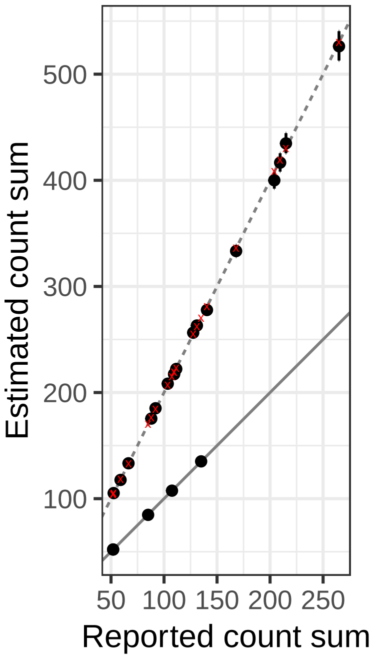

Figure 2: Estimated and reported chironomid count sums by the minumum percent method (solid symbols with error bars) and the direct search method (red crosses) from Last Chance Lake (Axford et al. 2017). Lines show the 1:1 (solid) and 2:1 (dashed) relationships.

Table 2: Caption

dataset	Number of assemblages	Median number of taxa	GCD	Median number of singletons (GCD =
1	53	36	0.5	
2	83	36	0.6	
3	27	19	0.7	
4	41	19	0.9	

(Fig. 3). The latter have a richness that is typically of the former with count sums half as large.

One assemblage with a greatest common divisor of two has 73 taxa. Assuming the distribution of counts is (at least locally) uniform, the probability of n counts being divisible by k is $1/k^n$: for a single assemblage with this many taxa, the probability is 1 in 10^{21}. The probability that these counts reflect the true nature of the pollen assemblages is exceedingly low; it is far more likely that the data have been mishandled at some stage in some way.

3.4. Unexpected percent data

Dataset diatom1 includes 35 diatom assemblages with a median richness of 19 taxa. The associated paper reports that at least 400 valves were counted from each assemblage, which means that singletons should have a relative abundance of 0.25% or less. However, the rarest taxon in two assemblages have a relative abundance of 3.2%, and the relative abundance of all other taxa in these assemblages are, within rounding error, integer multiples of this. If the count sums actually are at least 400, this would imply that each species count in these assemblages is an integer multiple of at least 13. This seems unlikely. Alternatively, the actual count sum for these assemblages could be as low as 31 valves. In total, six assemblages appear to have a count sum below 400 valves.

The archived data include 56 taxa that have a maximum abundance of at least 1.25%; this is about half of the 109 taxa reported in the paper. This pruning of rare taxa will increase the risk of counts without singletons, which would cause the minimum percent method to fail, but will have little impact on the direct search method.

Dataset diatom2 has 71 assemblages with a median richness of 57 taxa. The associated paper reports that the diatom counts included 400–500 valves per assemblage except for four diatom-poor assemblages with at least 100 valves. The direct search method estimates that twelve assemblages have a count sum of less than 400, and three have less than 100. The direct search and minimum percent methods agree, within rounding error, on the estimated count sums, implying that either all assemblages have singletons and count sums are low in some assemblages, or that the greatest common divisor of the raw counts is greater than one. This means that for the assemblage with the minimum percent is 7.1, which has seven taxa, all counts would need to be multiples of 8 for the count sum to be at least 100.

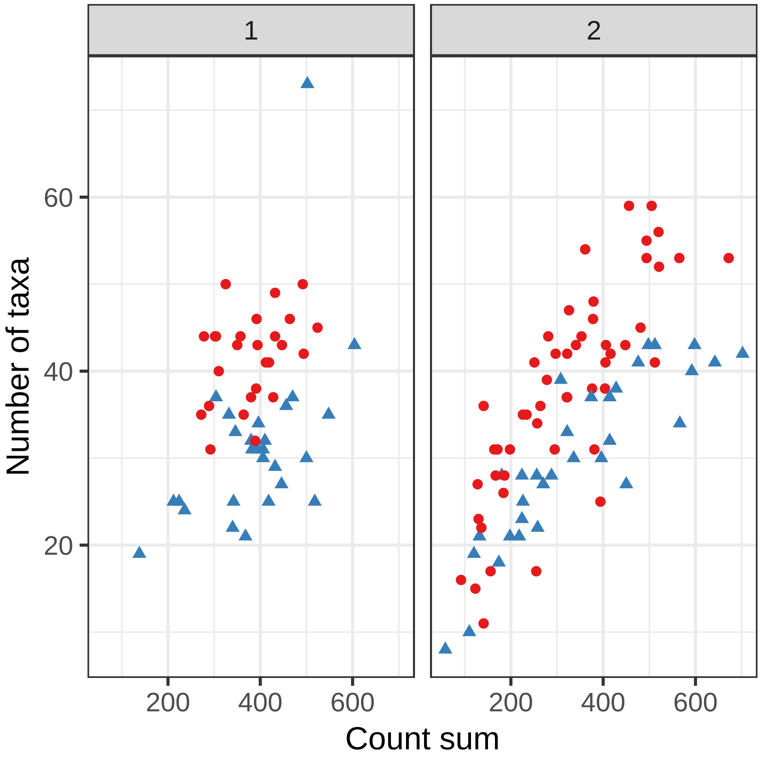

Figure 3: Number of taxa against count sum for assemblages with a greatest common divisor of one (red circles) or two (blue triangles). All four datasets show the same pattern, but it is clearest in datasets 1 and 2 which have the highest proportion of assemblages with a greatest common divisor of two.

Dataset chironomid1 has 55 assemblages. Although the associated paper reports that assemblages with count sums below 50 were discarded, the direct search method estimates that three assemblages had count sums between 38 and 40. These assemblages are diverse (richness between 15 and 17 taxa), so it is unlikely that the true counts are double that estimated here and the greatest common divisor is two.

Palaeoceanographic dataset marine1 has 160 assemblages with a median taxonomic richness of 34.5 taxa. The associated paper reports that assemblages with count sums below 100 were omitted, and that there were 20 assemblages with count sums below 200. The direct search method estimates that five assemblages have count sums below 100 (this excludes one mono-specific assemblage), with count sums as low as 34. With this dataset, the direct search and the minimum percent method give identical results for some assemblages but highly divergent estimates for others, with the direct search method sometimes giving estimates of several thousand, in one case over two orders of magnitude higher than the minimum percent method. Some of the divergence appears to be because some, mainly low diversity, assemblages lack singletons. The most extreme divergences appear to be data entry errors with incorrect rounding, perhaps during taxonomic revisions of the percent data rather than the raw counts.

4. Discussion

Analysis of the breeding bird, Owen's Lake, testate amoeba and pollen data sets has shown that, as expected, the vast majority of community and assemblages counts should have singletons, and even more should have a greatest common divisor of one. This means that these characteristics are potentially useful for identifying data where the count sum is misreported.

My search through archived microfossil data found several datasets where count sums are, or appear to be, smaller than that reported. Some of the count sums appear to be an order of magnitude below what was reported. The percent datasets examined are a convenience sample of available data that met the inclusion criteria. As such, this analysis cannot be used to accurately determine the prevalence of datasets with possible undercounts, but it appears that a non-negligible fraction of the literature is affected. It is possible in some cases that the assemblages with low counts were omitted or merged prior to analysis, but the number of assemblages reported was not updated. Care was taken to identify any such cases by, for example, examining stratigraphic diagrams to see if the low count assemblages were included. In some cases, the assemblages with low apparent counts can be seen in the published stratigraphic diagrams.

It is also possible that the apparently low count sums are a false positive. This can only happen in assemblages that have a greatest common divisor greater than one, which is very rare, especially for taxonomically diverse assemblages. If a dataset contains several taxonomically diverse assemblages that have a greatest common divisor of two (or expecially if greater than two) doubts should be raised about whether the counts sums are accurately described.

Microfossil percent data are often given to two decimal places. This is sufficient precision for the direct search method to have utility with counts sums of several thousand. Some data are only given to the nearest percent, which means that neither method has utility with count sums above 100.

The direct search method will fail if the dataset includes taxa calculated with different count sums, for example pollen sums of trees, shrubs and upland herbs and a pollen and spore sum that also includes pteridiophytes. It should be possible to identify such cases from the meta-data and knowledge of usual practice with different proxies. It will also fail, as show by dataset marine1, if percentages are incorrectly calculated or rounded.

Some taxonomic groups have microfossils that often come in groups of attached individuals. For example, for each diatom cell has two valves which are the counting unit. These valves usually separate during processing, but paired valves are often found, and some taxa such as *Aulacoseira* produce long chains of strongly bound valves. This might make singletons slighly less likely. The results from Owen's Lake suggest that this is not an important problem, as most assemblages have many singletons.

Small undercounts in a few assemblages will have minimal impact on the precision of any palaeoenvironmental reconstruction or other statistics derived from the assemblage. Substantial undercounts will potentially seriously effect the precision of the results. Such undercounts might constitute a data handling error, for example, if samples with low counts were supposed to be merged but that step was forgotten. Substantial and pervasive undercounting could be construed as scientific misconduct due to either negligence or falsification, and action to correct the literature is probably required.

Some papers do not report count sums. When this important quality metric is omitted, the reader should be able to assume that the standard minimum count sum for the taxonomic group has been used (i.e. 50 for chironomids, several hundred for pollen and diatoms). If the actual count sums are materially below this, then this potentially constitutes falsification by omission (Fanelli 2013).

5. Conclusions

A non-negligible fraction of the archived assemblage data includes counts that appear to have a count sum below that reported in the paper. Some of the apparent counts are small enough that the uncertainty of the count and derived statistics will be substantially larger than expected. These results highlight the importance of archiving both the raw data and the code required to process the data.

Acknowledgements

Some data were obtained from the Neotoma Paleoecology Database (http://www.neotomadb.org), and the work of the data contributors and the Neotoma community is gratefully acknowledged.

References

Amesbury MJ, Booth RK, Roland TP et al (2018) Towards a Holarctic synthesis of peatland testate amoeba ecology: Development of a new continental-scale palaeohydrological transfer function for North America and comparison to European data. Quaternary Science Reviews 201:483–500. doi: 10.1016/j.quascirev.2018.10.034

Axford Y, Levy LB, Kelly MA et al (2017) Timing and magnitude of early to middle Holocene warming in East Greenland inferred from chironomids. Boreas 46:678–687. doi: 10.1111/bor.12247

Barabesi L, Cerasa A, Cerioli A, Perrotta D (2018) Goodness-of-fit testing for the Newcomb-Benford Law with application to the detection of customs fraud. Journal of Business & Economic Statistics 36:346–358. doi: 10.1080/07350015.2016.1172014

Bik EM, Fang FC, Kullas AL et al (2018) Analysis and correction of inappropriate image duplication: The Molecular and Cellular Biology Experience. Molecular and Cellular Biology 38:e00309–18. doi: 10.1128/MCB.00309-18

Borchers HW (2018) Numbers: Number-theoretic functions

Bradbury JP (1997) A diatom-based paleohydrologic record of climate change for the past 800 k.y. from Owens Lake, California. In: An 800,000-year paleoclimatic record from core OL-92, Owens Lake, Southeast California. Geological Society of America

Brown NJL, Heathers JAJ (2017) The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. Social Psychological and Personality Science 8:363–369. doi: 10.1177/1948550616673876

Carlisle JB (2017) Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia 72:944–952. doi: 10.1111/anae.13938

Coddington JA, Agnarsson I, Miller JA et al (2009) Undersampling bias: The null hypothesis for singleton species in tropical arthropod surveys. Journal of Animal Ecology 78:573–584. doi: 10.1111/j.1365-2656.2009.01525.x

Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London

Fanelli D (2013) Redefine misconduct as distorted reporting. Nature 494:

Grieneisen ML, Zhang M (2012) A comprehensive survey of retracted articles from the scholarly literature. PLOS ONE 7:1–15. doi: 10.1371/journal.pone.0044118

Heiri O, Lotter AF (2001) Effect of low count sums on quantitative environmental reconstructions: an example using subfossil chironomids. Journal of Paleolimnology 26:343–350. doi: 10.1023/a:1017568913302

Larocque-Tobler I, Filipiak J, Tylmann W et al (2015) Comparison between chironomid-inferred mean-August temperature from varved Lake Żabińskie (Poland) and instrumental data since 1896 AD. Quaternary Science Reviews 111:35–50. doi: 10.1016/j.quascirev.2015.01.001

Larocque-Tobler I, Filipiak J, Tylmann W et al (2016) Corrigendum to "Comparison between chironomid-inferred mean-August temperature from varved Lake Żabińskie (Poland) and instrumental data since 1896 AD" [Quat. Sci.

Rev. 111 (2015) 35–50]. Quaternary Science Reviews 140:163–167. doi: 10.1016/j.quascirev.2016.01.020

Pardieck K, Ziolkowski D, Lutmerding M, Hudson M-A (2018) North American breeding bird survey dataset 1966 - 2017, version 2017.0. https://doi.org/10.5066/F76972V8

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Telford RJ (2019a) Review and test of reproducibility of sub-decadal resolution palaeoenvironmental reconstructions from microfossil assemblages

Telford RJ (2019b) CountSum: Check assemblage count sums and percent

Williams JW, Grimm EC, Blois JL et al (2018) The neotoma paleoecology database, a multiproxy, international, community-curated data resource. Quaternary Research 89:156–177. doi: 10.1017/qua.2017.105

Wolodzko T (2018) ExtraDistr: Additional univariate and multivariate distributions