13.2 Analysing cumulative standardised data

Rationale

Suppose we have data where the variables consist of different size classes of mussels (as we shall shortly see in a real example). In such cases, where the *variables have a natural order*, we may, of course, simply treat the variables multivariately just as they are, ignoring their intrinsic quantitative inter-relationships and natural ordering. However, note that Bray-Curtis (or Manhattan, or whatever measure we apply) will take no notice of the ordering of the variables. In other words, we could shuffle the order of the variables and it would make no difference at all to the calculated resemblance matrix among samples. However, we may want to permit the natural ordering of the variables to play a useful and meaningful role in the analysis itself.

If we use *cumulative percentage values* across each sample, then we can really see the sample as a 'profile' of size classes, from smallest to largest. This is different from viewing the sample as a set of individual size-class variables that bear no relationships to one another (i.e., where distances between those size classes would be of no importance). By standardising each sample to a *cumulative profile across the size classes*, we acknowledge that the variables are, themselves, structured, from smallest to largest; i.e., they are ordered. Of course, the variables will not be independent of one another following a standardisation like this, but we do not, typically, expect the variables to be independent of one another in multivariate analyses to begin with. It is the simultaneous action of all variables (in this case, the overall *shape of the profile*) that is relevant to us for comparative analysis among the samples.

Size-class data: a 'toy' example

To make this more concrete, consider the following 'toy' example dataset, where we have 5 samples (columns), and each of these have the following *raw percentages* of individuals of a given species belonging to each of 8 size classes (rows, which are the variables here) and these are ordered.

Toy example

Size class data - raw percentages

Abundance

Variables	Samples				
	1.More.small	2.Some.small	3.Even	4.Some.large	5.More.large
2-4 mm	50	0	12.5	0	0
4-6 mm	0	50	12.5	0	0
6-8 mm	50	0	12.5	0	0
8-10 mm	0	50	12.5	0	0
10-12 mm	0	0	12.5	50	0
12-14 mm	0	0	12.5	0	50
14-16 mm	0	0	12.5	50	0
16-18 mm	0	0	12.5	0	50

The sample labelled '3.Even' has a perfectly even distribution, with 12.5% of the individuals in every size class. The sample labelled '1.More.small' has 50% of its individuals in the smallest size-class (2-4 mm), and 50% in the 6-8 mm size-class. The sample labelled '2.Some.small' has 50% of its individuals in the 4-6 mm size-class, and 50% in the 8-10 mm size-class, and so on.

Now, the Manhattan distances between all pairs of samples here looks like this:

Resem1

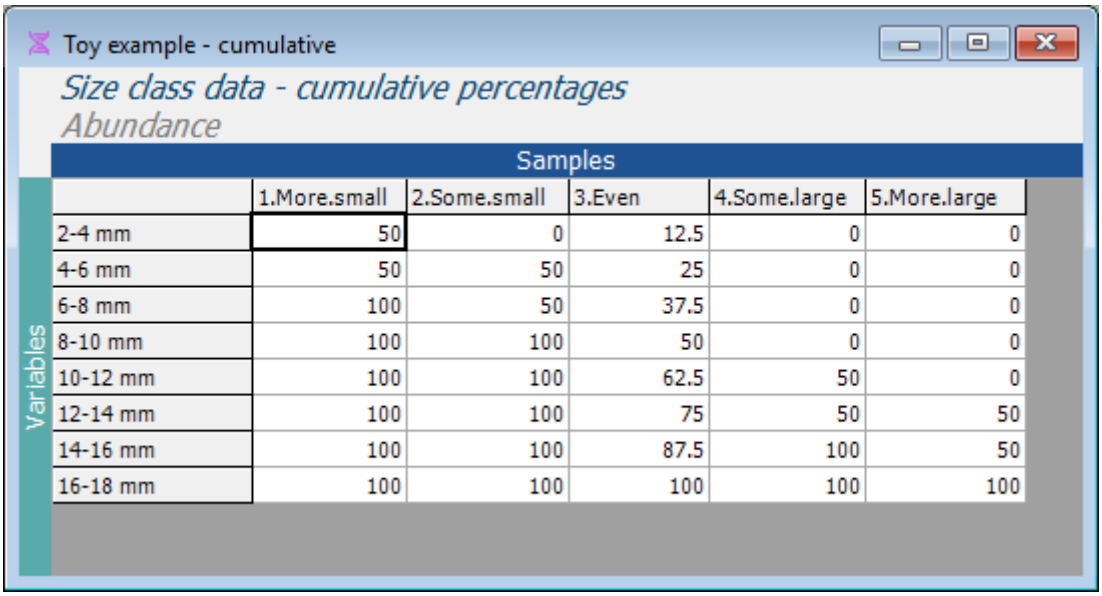
Size class data - raw percentages

Distance (0 to inf)

Samples	Samples				
	1.More.small	2.Some.small	3.Even	4.Some.large	5.More.large
1.More.small					
2.Some.small	200				
3.Even	150	150			
4.Some.large	200	200	150		
5.More.large	200	200	150	200	

Note that all of the samples are an equal distance away from the '3.Even' sample (150 units), and all other pairs of samples are considered to be 200 units away from each other, which is the maximum possible distance we could get using the Manhattan measure on these data. This is despite the fact that, biologically, we would rather prefer '2.Some.small' to be closer to '1.More.small' than it is to (say) '4.Some.large' or '5.More.large'. You can see that the above analysis completely ignores the fact that the variables themselves are ordered.

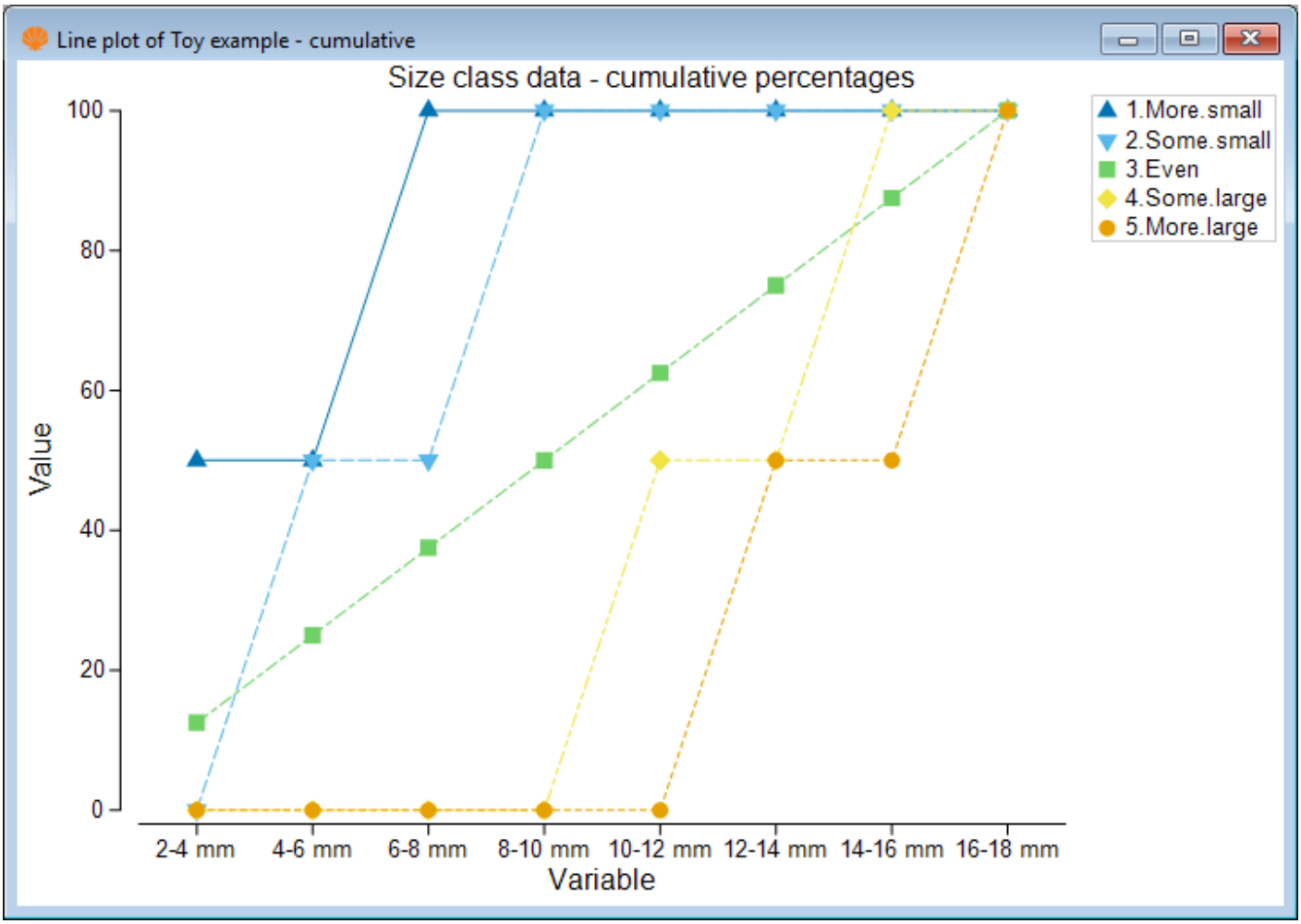
Now, let’s standardise the above raw percentages to a cumulative profile of percentages. Here, we cumulatively add the percentages in each size class across the sample so that, by the end of the list, we arrive at 100%. The cumulative percentages for this toy example look like this:



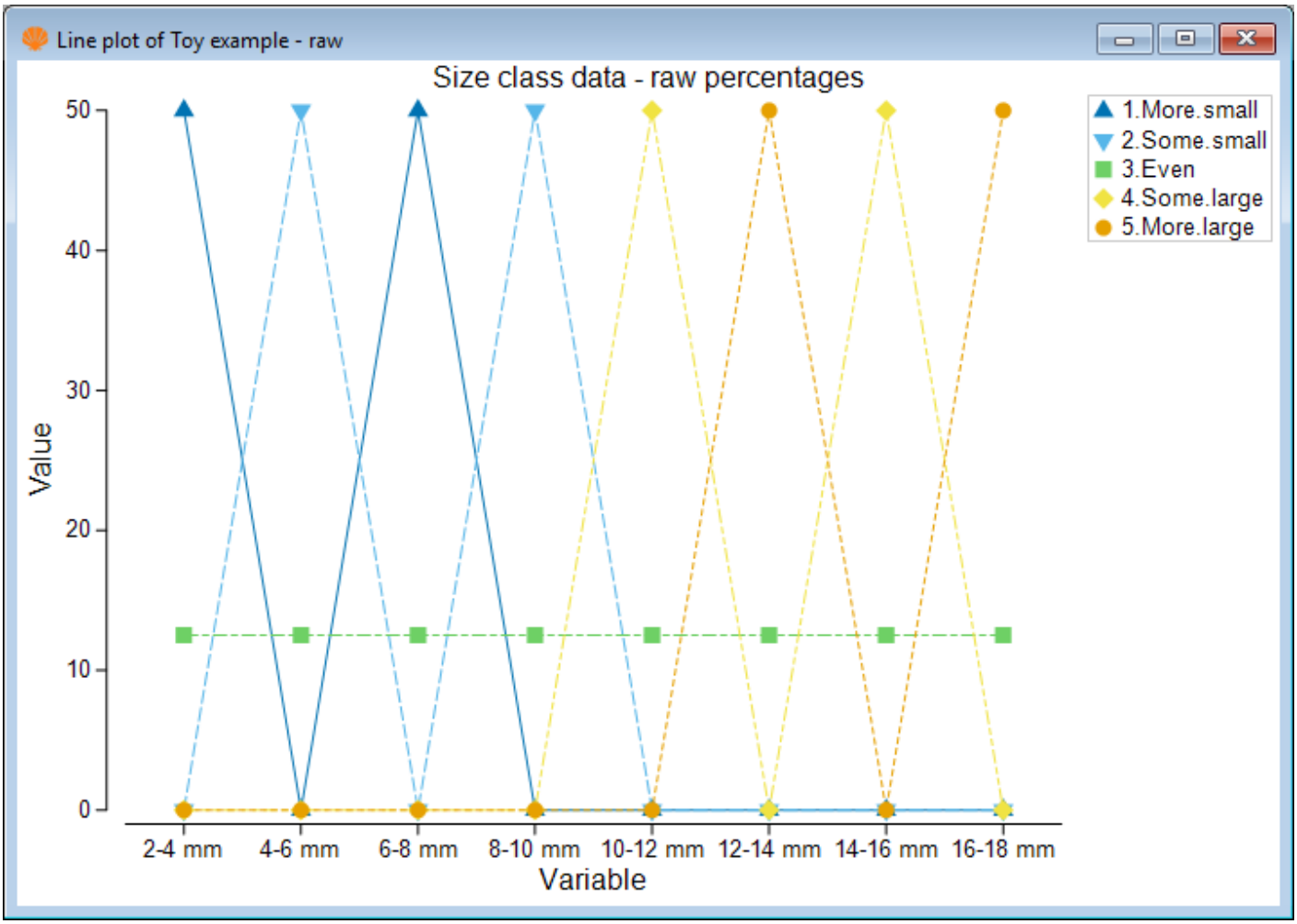
	1.More.small	2.Some.small	3.Even	4.Some.large	5.More.large
2-4 mm	50	0	12.5	0	0
4-6 mm	50	50	25	0	0
6-8 mm	100	50	37.5	0	0
8-10 mm	100	100	50	0	0
10-12 mm	100	100	62.5	50	0
12-14 mm	100	100	75	50	50
14-16 mm	100	100	87.5	100	50
16-18 mm	100	100	100	100	100

Let's think about what this transformation to cumulative percentages has done. Consider the '3.Even' column first. We start with 12.5% of the individuals in the 2-4 mm size class, then we add another 12.5% at the next step in the profile, 4-6 mm, which gets us to 25%, then we add another 12.5% at the next step in the profile for the 6-8 mm size class, which gets us to 37.5%, and so on.

Visually, we have gone from a series of unrelated (unordered) numbers to a profile of numbers which increases from left to right along the size classes, from 0% to 100% of the sample. This is perhaps best visualised using a line plot¶, where the size classes are along the x-axis and the y-axis has the cumulative percentage values (see below). Of course, by the time we get to the final size class, we end up at 100% (and, for any given sample, we may well end up at 100% far before we reach the final size class, which is fine).

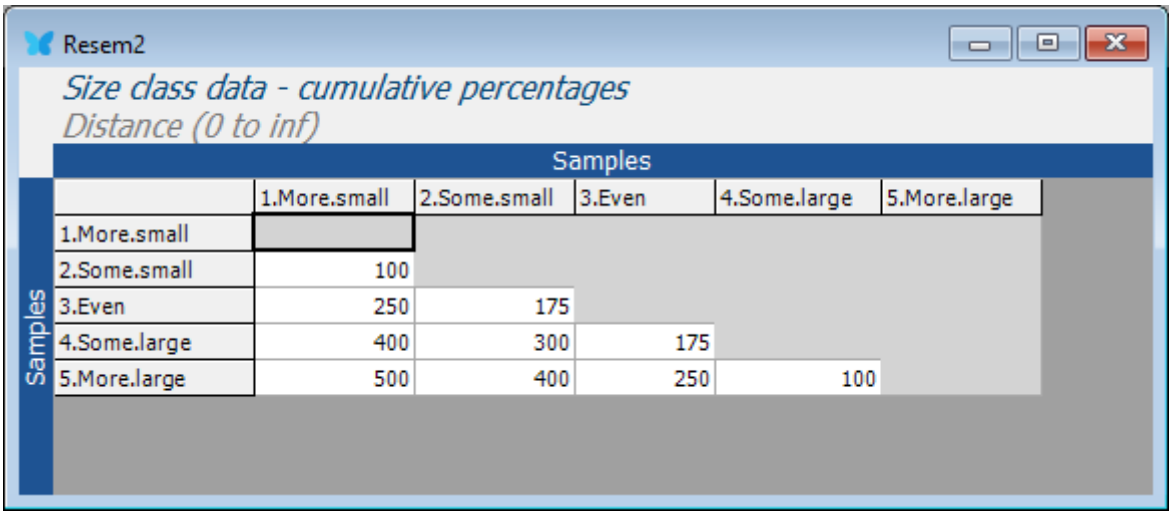


Now, in the above plot, we can see that the evenly distributed sample (in green) just marches along at equal step lengths from left to right, while the two samples that contain a large percentage of small individuals (in dark blue and light blue) reach 100% rather quickly, whereas the samples that contain larger individuals (in yellow and orange) do not increase towards 100% until the larger size classes are reached. This contrasts sharply with what the image looks like if we just plot the raw percentages:



The plot above is a bit more of a mess, and it ignores the important additional information we have about the size-class variables themselves; namely, that these variables are ordered!

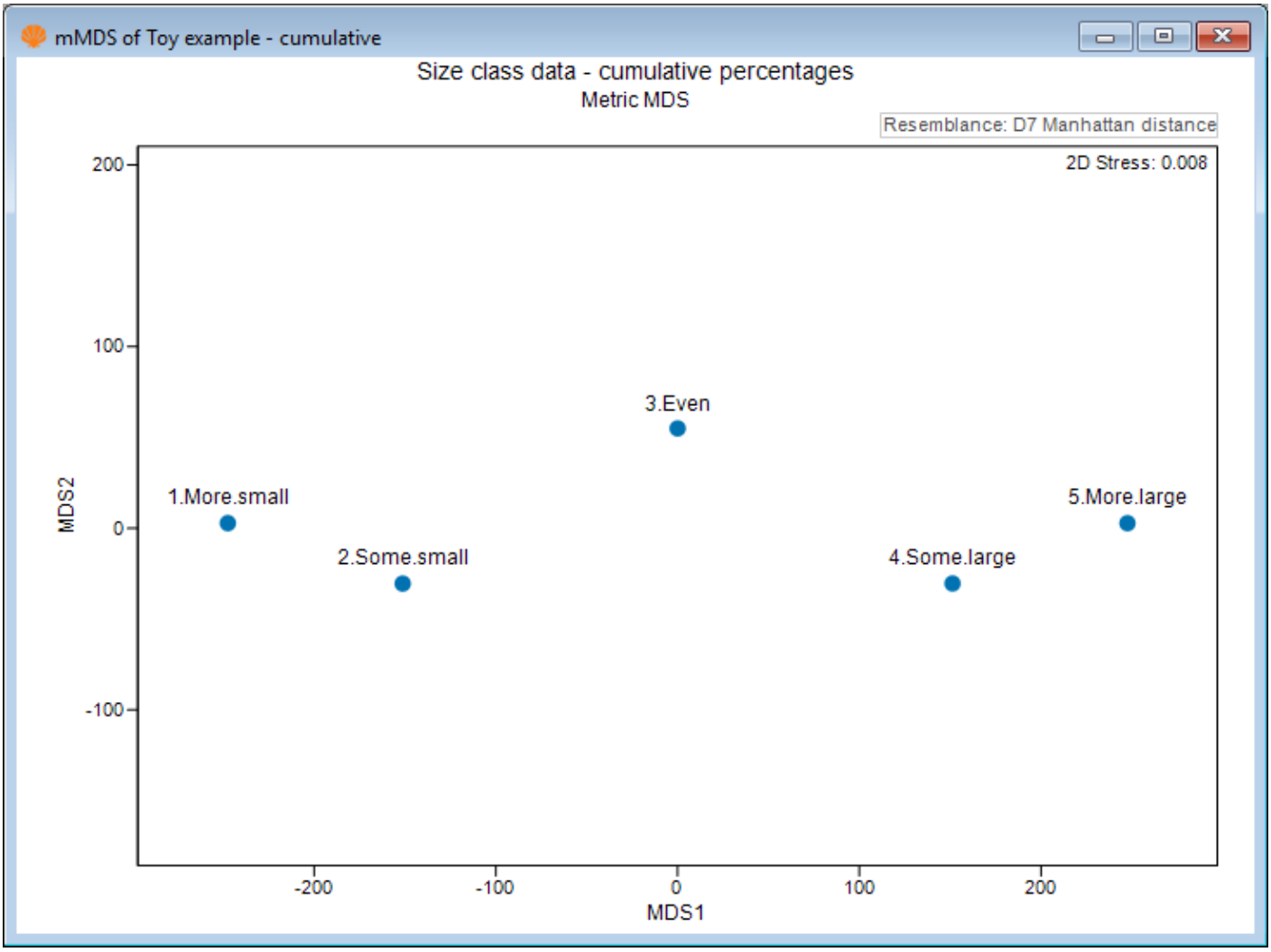
Next, if you calculate the Manhattan distances among each pair of samples using the *cumulative* percentages, it makes a lot more sense (see below).



	1.More.small	2.Some.small	3.Even	4.Some.large	5.More.large
1.More.small					
2.Some.small	100				
3.Even	250	175			
4.Some.large	400	300	175		
5.More.large	500	400	250	100	

For example, the largest distance (500 Manhattan units) is between the profile of '1.More.small' and '5.More.large'. Also, the distance between '1.More.small' and '2.Some.small' is less than that between '1.More.small' and either of '4.Some.large' or '5.More.large', and so on.

For even greater clarity, we can examine the metric MDS plot that we obtain based on the Manhattan distances calculated from raw percentages versus the one obtained from cumulative percentages (shown below).



Of course, the one using raw percentages really makes no sense at all, whereas the one based on cumulative percentages makes perfect sense!

Summary

Generally, if we are dealing with variables that are ordered, it make sense to treat them in a cumulative fashion and to analyse profiles, as demonstrated above. Another possibility would be to incorporate distances among variables in the calculation of the resemblances among samples. An example of this is taxonomic dissimilarity (Clarke *et al.* (2006b)), which includes the taxonomic or phylogenetic relationships among species in the calculation of resemblances among samples. One can, however, use any distance/dissimilarity matrix among the variables (whether they be species or not) within such a calculation. For example, Myers *et al.* (2021) used this approach to calculate functional dissimilarities among fish assemblages along depth and latitude gradients.

¶*In PRIMER 8, one has the flexibility to draw line plots either: (i) of samples (across variables, as done above) or (ii) of variables (across samples).*

Revision #30
Created 6 July 2025 22:23:16 by Marti
Updated 5 December 2025 21:41:11 by Marti