15. Other new tools & utilities

15.1 New default colour palette

Accessibility

It is important to make graphics accessible to those with color vision deficiencies. We have therefore re-vamped the colour palette for P8 to achieve distinctive colours for plots (by default) that carefully accommodate the most common forms of colour blindness (Fig. 15.1).

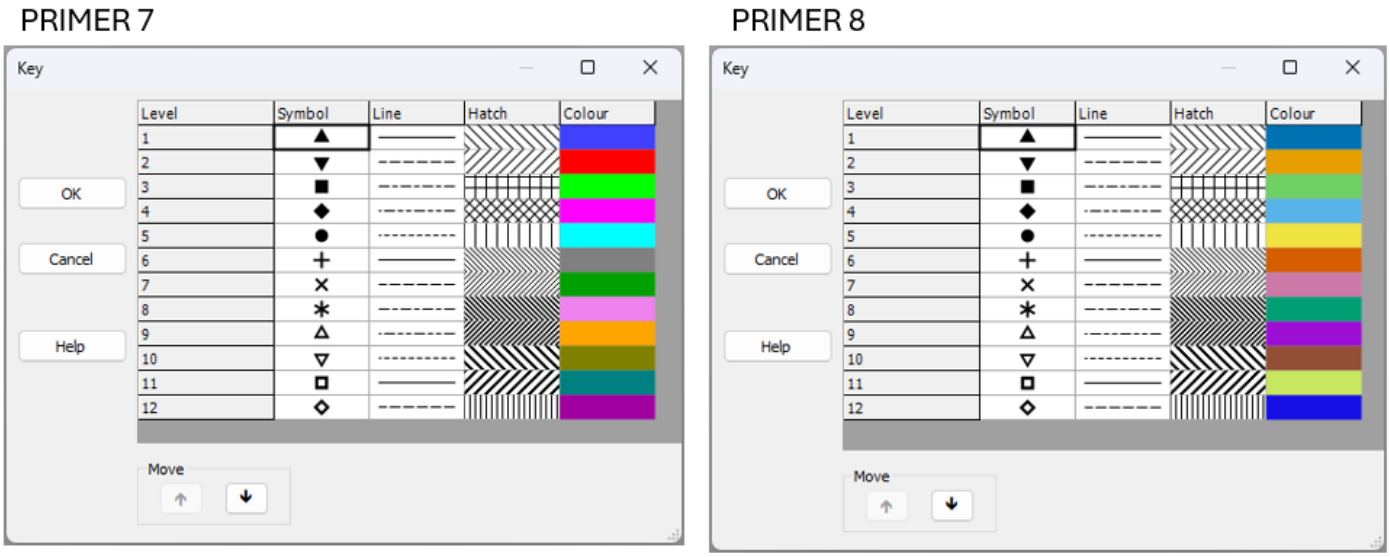


Fig. 15.1 New default colour palette in PRIMER 8 (at right) compared to the old default colour palette in PRIMER 7 (at left).

In developing these new defaults, we used the following resources:

- "Coloring for Colorblindness" - David Nichols
- "Points of view: Color blindness" - Bang Wong
- Paul Tol's Notes: Colour schemes and templates - Paul Tol

We especially appreciated David Nichols' online tool, which allowed us to play with colour schemes and get a feel for how different schemes would look for people who experience colour differently, due to various forms of alterations to their color vision, including protanopia, deuteranopia, tritanopia or deuteranomaly. We ended up with the following pallette:

- PRIMER 8 default colours - (see the left-hand column labeled 'True' when you follow this link).

To get there, we borrowed the base colour scheme from Wong (2011) . We then simply re-arranged the order of the colours and extended the scheme to 12 colours which we felt, as much as possible, would still appear well-differentiated from one another, right across the range of

different forms of colour vision deficiency.

Compatibility with earlier versions of PRIMER

If you have an existing *.pri or *.pwk file that has been opened and saved in PRIMER 7, then it will carry around the default colour palette from PRIMER 7 with it. Thus, when you open up a PRIMER 7 file in PRIMER 8, those old default colours from PRIMER 7 (or whatever other colours you tweaked and saved in your file's graphics) will be used. To get rid of all of the older colours, and hence see the new default colour palette in P8 when you create a new graphic, please save the file as an Excel file and then import the data, fresh, into PRIMER 8.

15.2 New selection options

In PRIMER 8, the options available for *selecting samples* or *selecting variables* have been expanded considerably from what they were in PRIMER 7.

Selecting Samples

When you click **Select** > **Samples...** in P8, you will see the new dialog shown at right in Fig. 15.2.

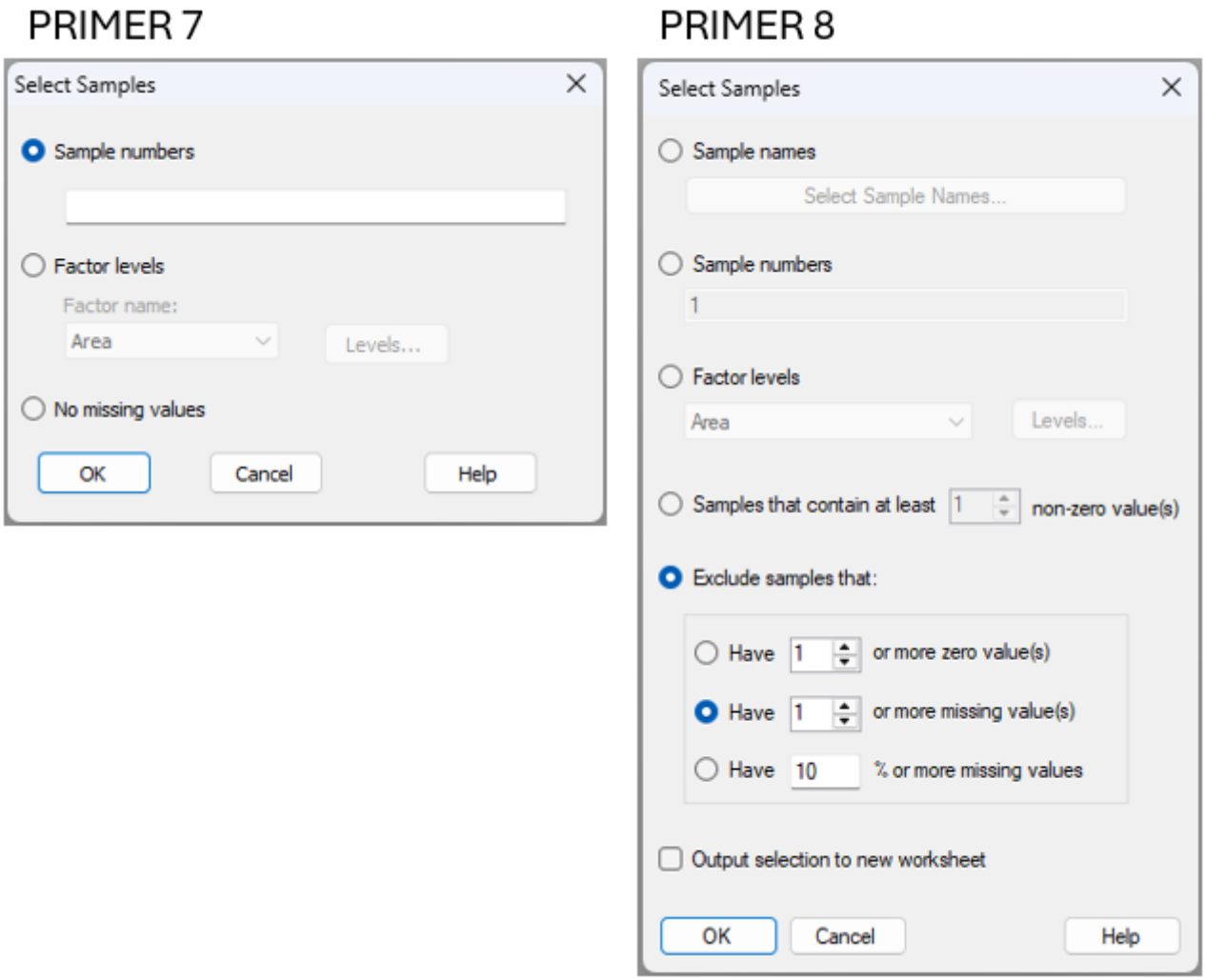


Fig. 15.2 Comparison of the Select > Samples... dialog window in PRIMER 7 (at left) vs PRIMER 8 (at right).

Note that in PRIMER 8, you can select samples:

- using **sample names** (this option includes a filter to help find names in long lists);
- using **sample numbers**;
- belonging to certain **level(s) of a factor**; or
- that contain some minimum number (x) of **non-zero values**.

You can also *exclude* samples that:

- have (x) or more zero values;
- have (x) or more missing values; or
- have (x)% or more missing values.

Note also that there is the option to: '$\checkmark$Output selection to new worksheet'.

This latter option means you don't have to take any extra steps (post-selection) simply to duplicate the selected data into a new worksheet (if desired) for subsequent analyses. As an added bonus, a small output file is produced when you tick this option which identifies precisely what choices you made to select the samples. It is very useful to have this information going forward, to keep track of any subset selections made along your analysis pathway.

Selecting Variables

When you click **Select** > **Variables...** in P8, you will see the new dialog shown at right in Fig. 15.3.

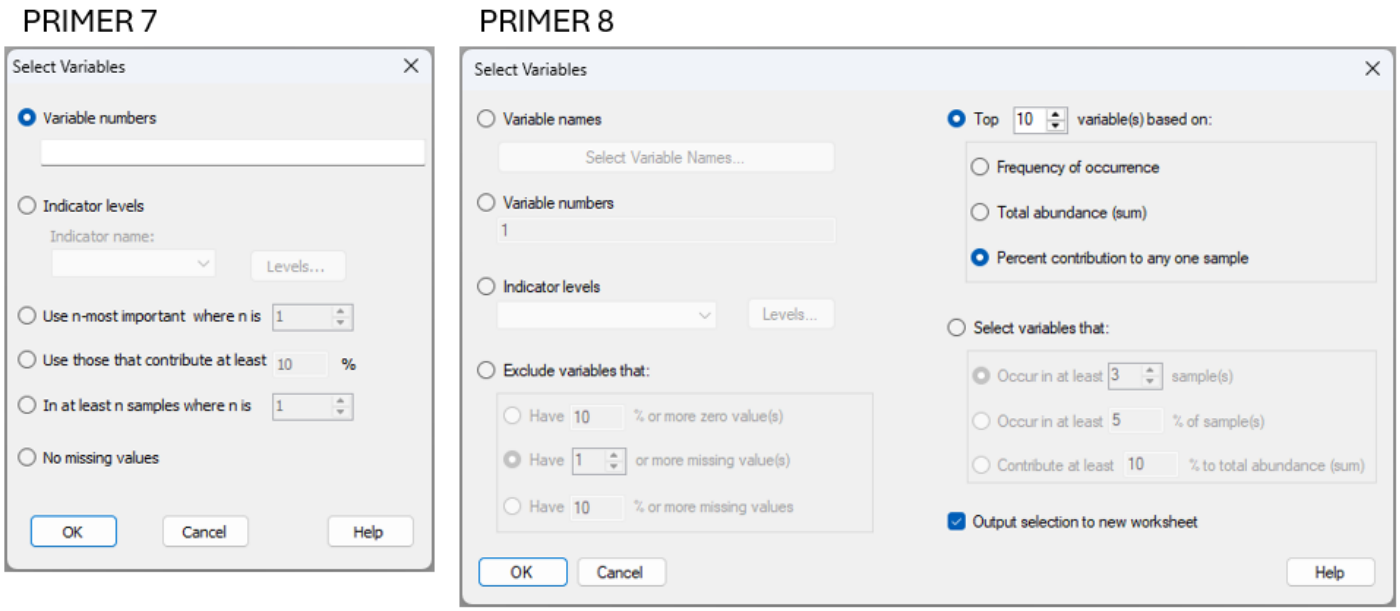


Fig. 15.3 Comparison of the Select > Variables... dialog window in PRIMER 7 (at left) vs PRIMER 8 (at right).

Note that in PRIMER 8, you can select variables:

- using **variable names** (this option includes a filter to help find names in long lists);
- using **variable numbers**; or
- belonging to certain **group(s) of an indicator**.

It is also possible to select variables that correspond to the top (x) variables in a list based on:

- **frequency of occurrence**;
- **total abundance** (sum); or
- **percent contribution to any one sample** (this option was called 'most important' in PRIMER 7)

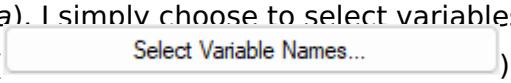
Finally, you may choose to select variables that:

- occur in at least (x) sample(s);
- occur in at least (x)% of sample(s); or
- contribute at least (x)% to the total abundance (sum).

Note that the above dialog gives you a lot of freedom with respect to identification of 'important' variables in ecological contexts, not just on the basis of total abundance (in any one sample or overall), but alternatively by reference to their frequency of occurrence. Using the frequency of occurrence may often be more suitable, as a criterion, than pulling out species based on their total abundance values, particularly if there are highly sporadic species that have massive abundance values (e.g., weeds or opportunists), but that may not actually be that important ecologically (e.g., they might have shown up in just a single sample).

As an added bonus (and precisely as we saw in the dialog for the selection of samples, shown above), you can choose to: '$\checkmark$Output selection to new worksheet'. This streamlines and clarifies analytical pathways.

Filters *(New!)*

For either the selection of samples or the selection of variables **by name**, PRIMER 8 offers a new tool in the form of a *filter* that is very helpful whenever you are dealing with long lists of names in large data sheets. For example, suppose I wish to select all of the variables from a long list of species that belong to the same genus (e.g., *Ampelisca*). I simply choose to select variables by names and click the 'Select Variable Names...' button (

), then start typing the genus name into the 'Filter:' box, and those species will appear in the 'Available:' list for me to easily select them (see Fig. 15.4 below).

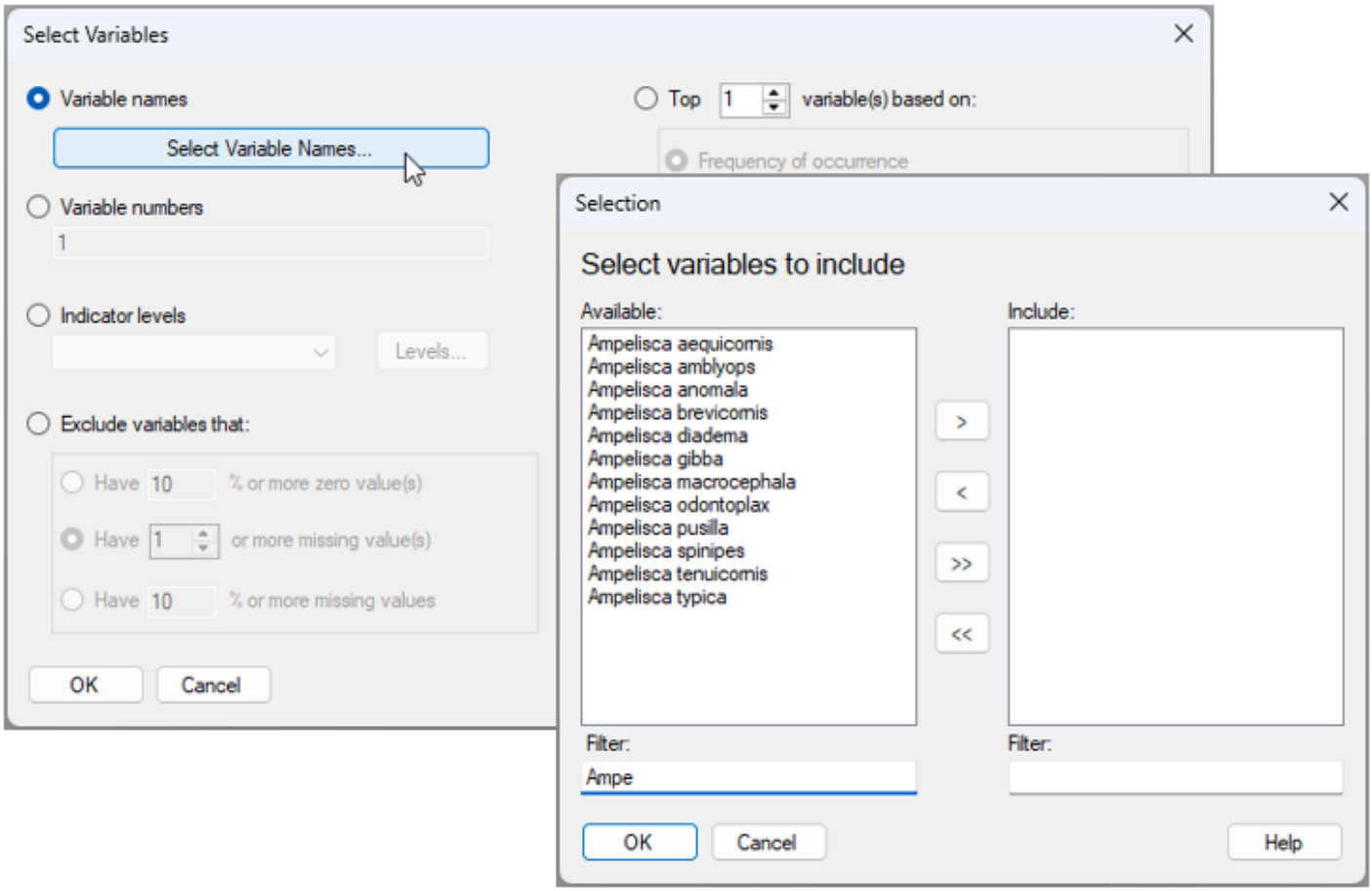


Fig. 15.4 Dialog window for selecting names of variables, using the new 'Filter:' tool in PRIMER 8. In the above example (data in the file named 'Norway_macrofauna.pri', found inside the 'Examples_P8 > Norway_macrofauna' folder), there are 809 taxa in the data sheet, but by typing the letters 'Ampe' in the 'Filter:' box, the available list is reduced to just the 12 variables that have this particular combination of letters in their name. All of these are of the genus 'Ampelisca'.

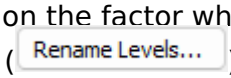
Using this new filtering tool in PRIMER 8 means you can quickly and easily find and select specific variables (or samples) that you know are in your data somewhere (e.g., all variables starting wtih the letter "R"), even if you cannot easily remember the entire name in detail.

15.3 Re-name levels of a factor (or indicator)

There are many situations where it would be very handy to be able to change the names of levels of a factor (or to change the names of groups for an indicator). We often need to tweak the names of levels of factors. For example,

- The names are too long and you want to shorten/abbreviate them so that they are easy to see as labels on an ordination plot. For example, you have levels of 'Berghan Point' and 'Home Point', and you want to change them to 'B' and 'H'.
- The factor has levels that are quantitative (e.g., such as depth zones), but the names are not strictly numeric (e.g., maybe the levels are named '50-100 m', '100-200 m', '200-300 m', etc.). You might, however, really want numeric factor level names so you can analyse the factor as **ordered** in an ANOSIM, put depth trajectories on ordination plots, etc.
- You have discovered that not all factor levels matter, and so you want to combine 2 or more factor levels into an amalgamated (single) factor level. For example, maybe you originally have levels of 'Low', 'Medimum-Low', 'Medium', and 'High', but later decide that the first two levels are not actually significantly different, so you want to combine those two and change the names so that you just have 'L', 'M' and 'H'.
- You have taken data through time (say, every 2 months) at your sites, and your labels for this temporal factor look like this: 'Jan-2001', 'Mar-2001', 'May-2001', etc., but you want to treat these simply as sequential time-points in your plots, e.g., 'T1', T2', 'T3', etc.

It would be very laborious to have to go through the process of re-naming the factor-level names for every single sample in the full worksheet, and this process would also be highly prone to error.

In PRIMER 8, all you need to do is click **Edit** > **Factors**, click on the factor whose level names you want to change, and then click the 'Rename Levels...' button (

). The resulting dialog lets you nominate the new names for the factor levels (just type them in), and then you can get on with your work.

For example, for the size-class data for mussels from several sites in the Gulf of Alaska (the data file is 'Gulf_of_Alaska_mussels.pri', located in the folder 'Examples_P8' > 'Gulf_of_Alaska_mussels'), we may wish to abbreviate the names of the ecoregions so that they are shorter, making them easier to see in ordination plots (Fig. 15.5).

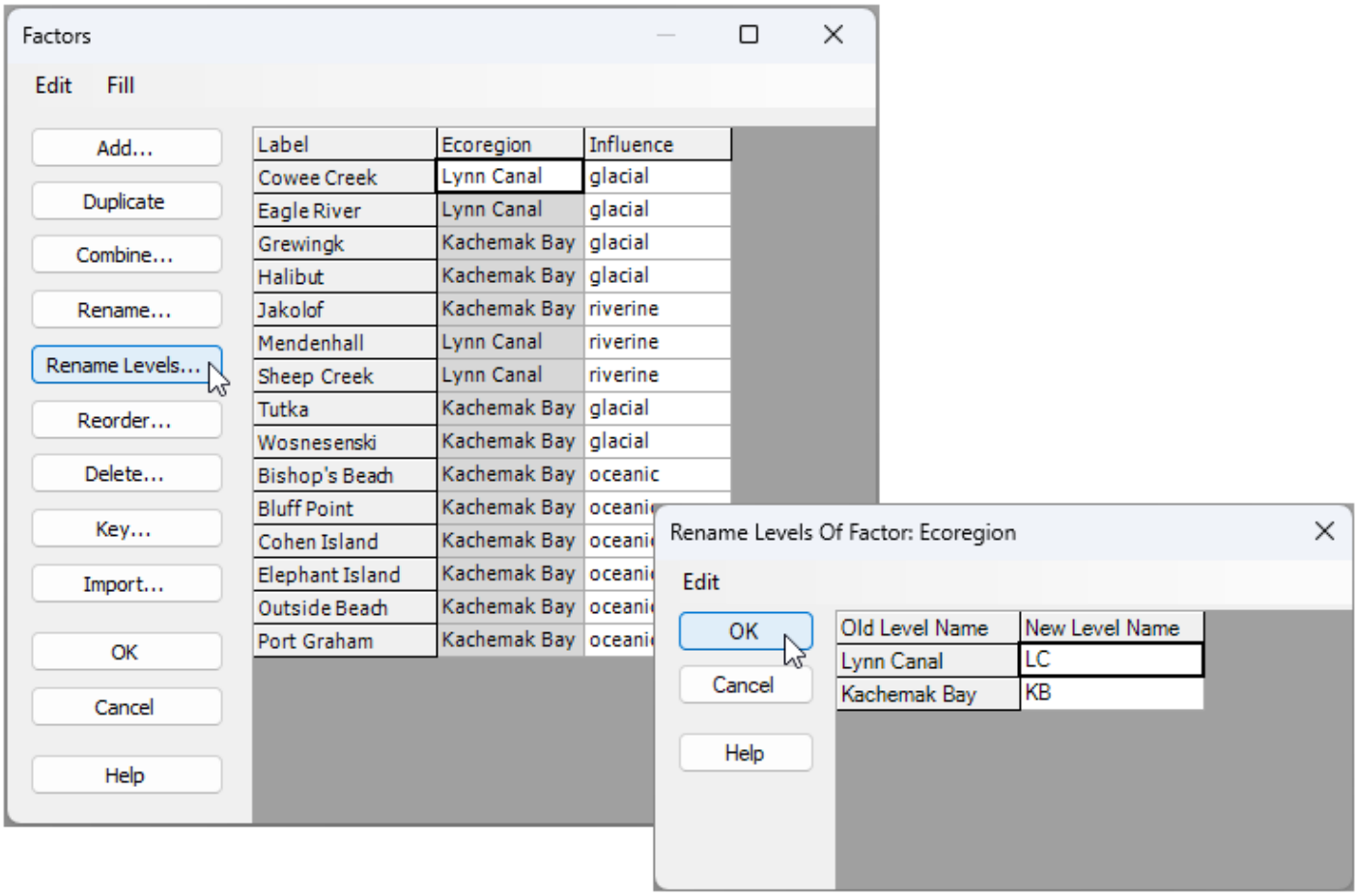
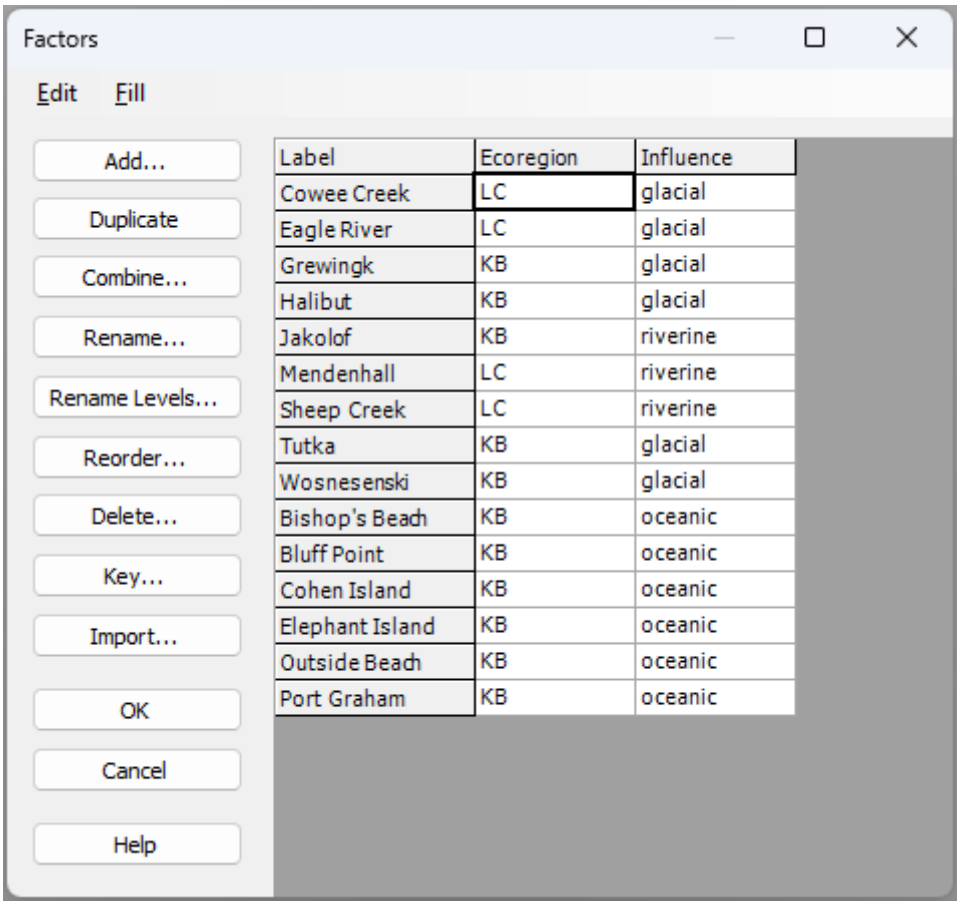


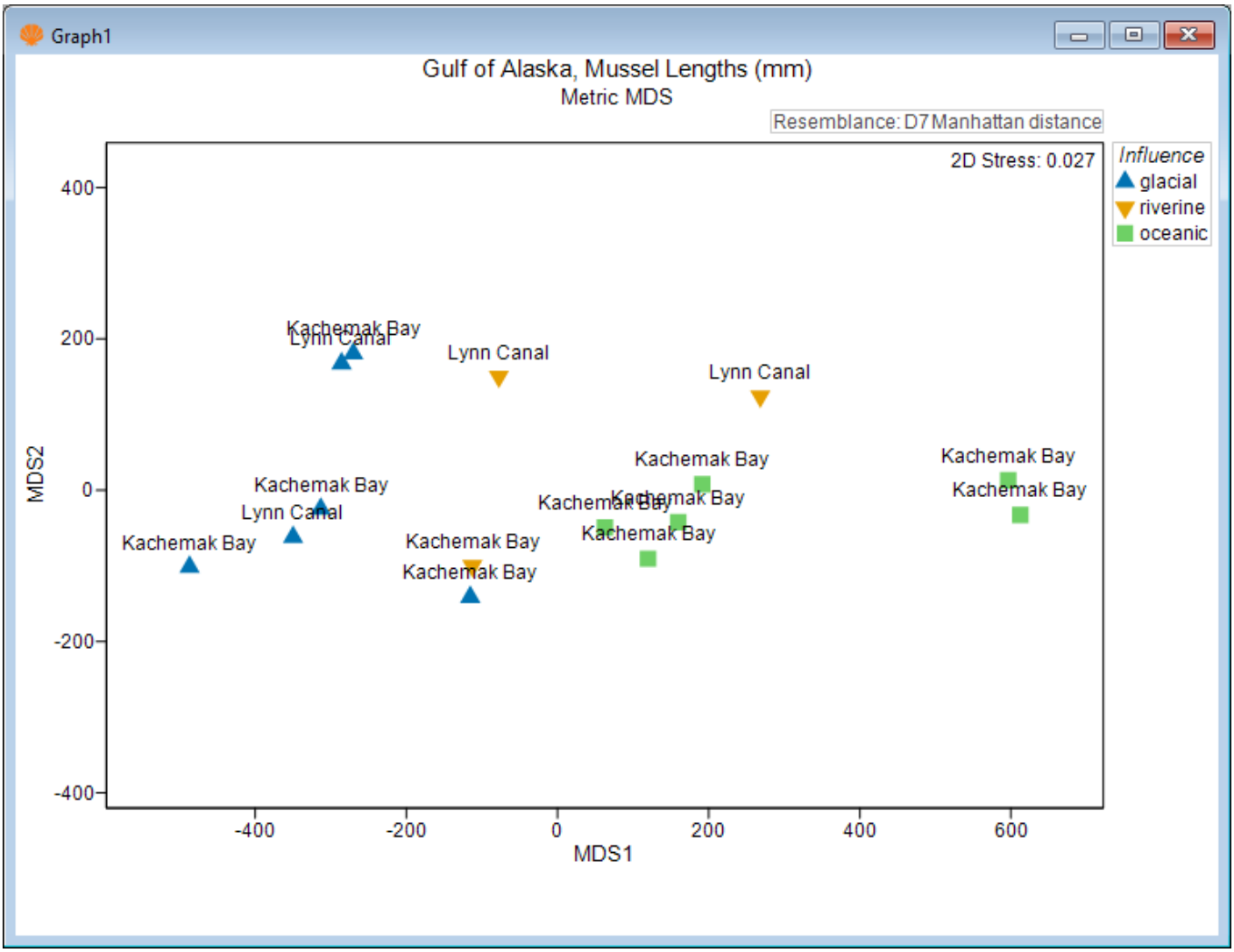
*Fig. 15.5 Dialog windows showing the action of changing the names of levels for the factor of '*Ecoregions*' for the Gulf of Alaska mussel size-class dataset: from '*Lynn Canal*' to '*LC*' and from '*Kachemak Bay*' to '*KB*'.*

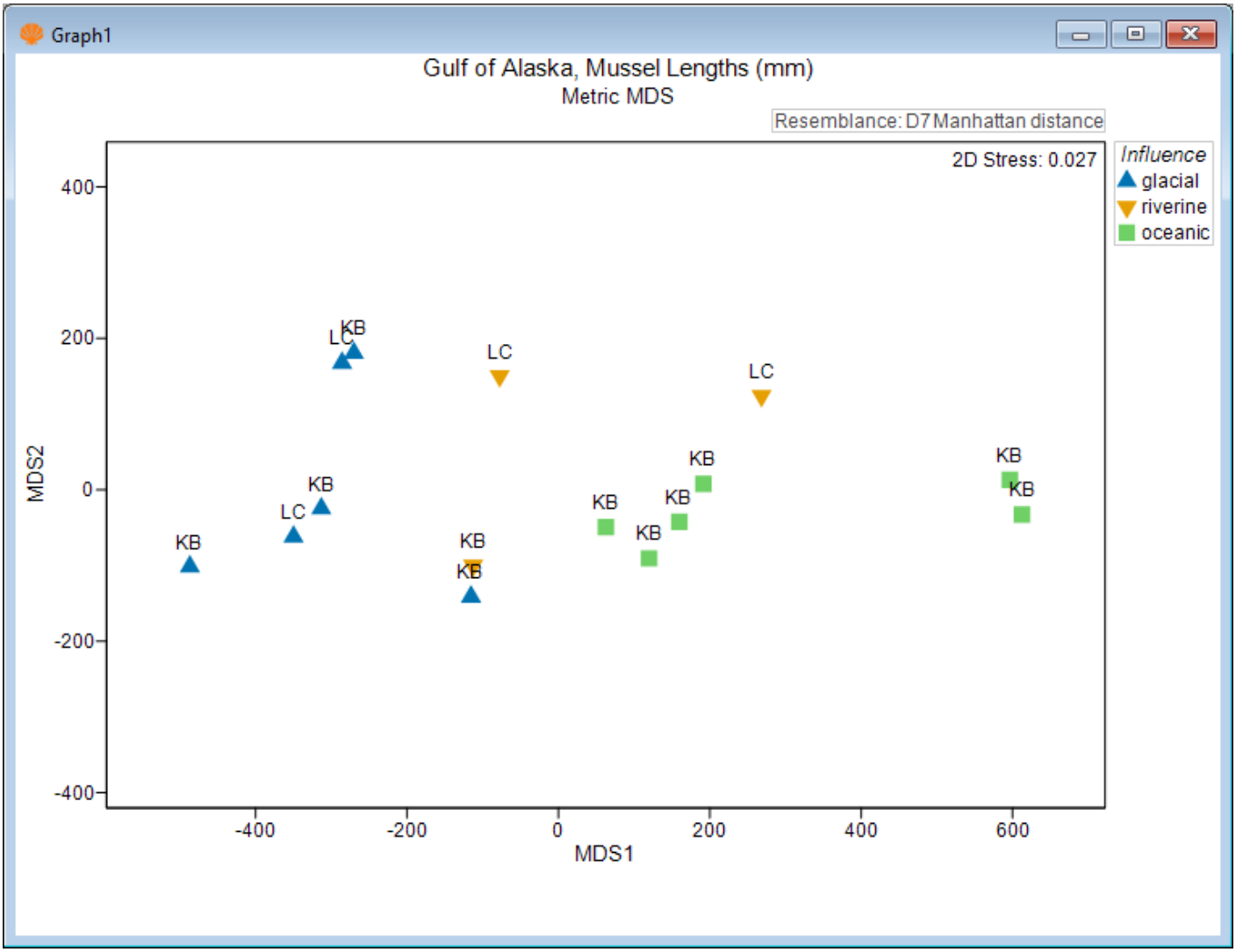
The resulting factor information (after changing the level names for 'Ecoregion', as shown above) looks like this:



As an aside, if you wanted to retain the original level names in your data sheet as well, then you need first to *duplicate* the original factor (using the 'Duplicate' button), *rename* the factor itself (using the 'Rename...' button; for example, the abbreviated names could be held in a factor called 'Ecoreg'), *then* create new level names for that duplicated factor (using the 'Rename Levels...' button) to whatever new names you wish.

You can compare the metric MDS plot of the sites (based on Manhattan distances among cumulative percentages of standardised size-classes of mussels) when Ecoregion factor-level labels have long names *vs* the same plot with short names (see below).


Graph1
Gulf of Alaska, Mussel Lengths (mm)
Metric MDS
Resemblance: D7 Manhattan distance
2D Stress: 0.027
Influence
glacial
riverine
oceanic
MDS2
MDS1
400
200
0
-200
-400
-400
-200
0
200
400
600
Kachemak Bay
Lynn Canal
Lynn Canal
Lynn Canal
Kachemak Bay
Kachemak Bay
Kachemak Bay
Kachemak Bay
Kachemak Bay
Kachemak Bay
Kachemak Bay
Kachemak Bay
Kachemak Bay
Kachemak Bay

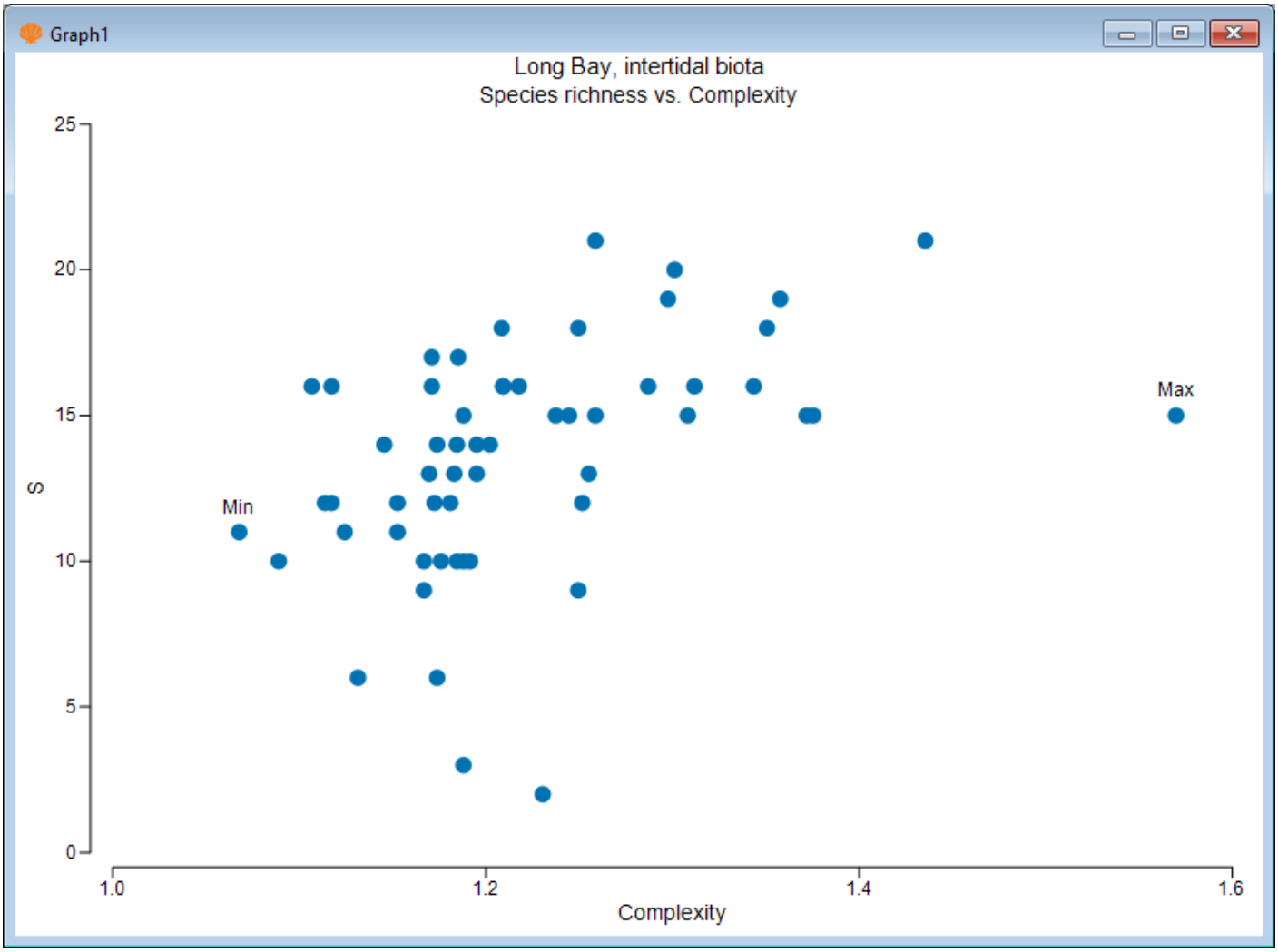


Finally, note that this same re-naming tool is also available under **Edit** > **Indicators...**.

15.4 Add customised values/labels to graphical axes

In PRIMER 8, there is a new tool that allows us to *add customised values and labels* to coordinate axes in graphics.

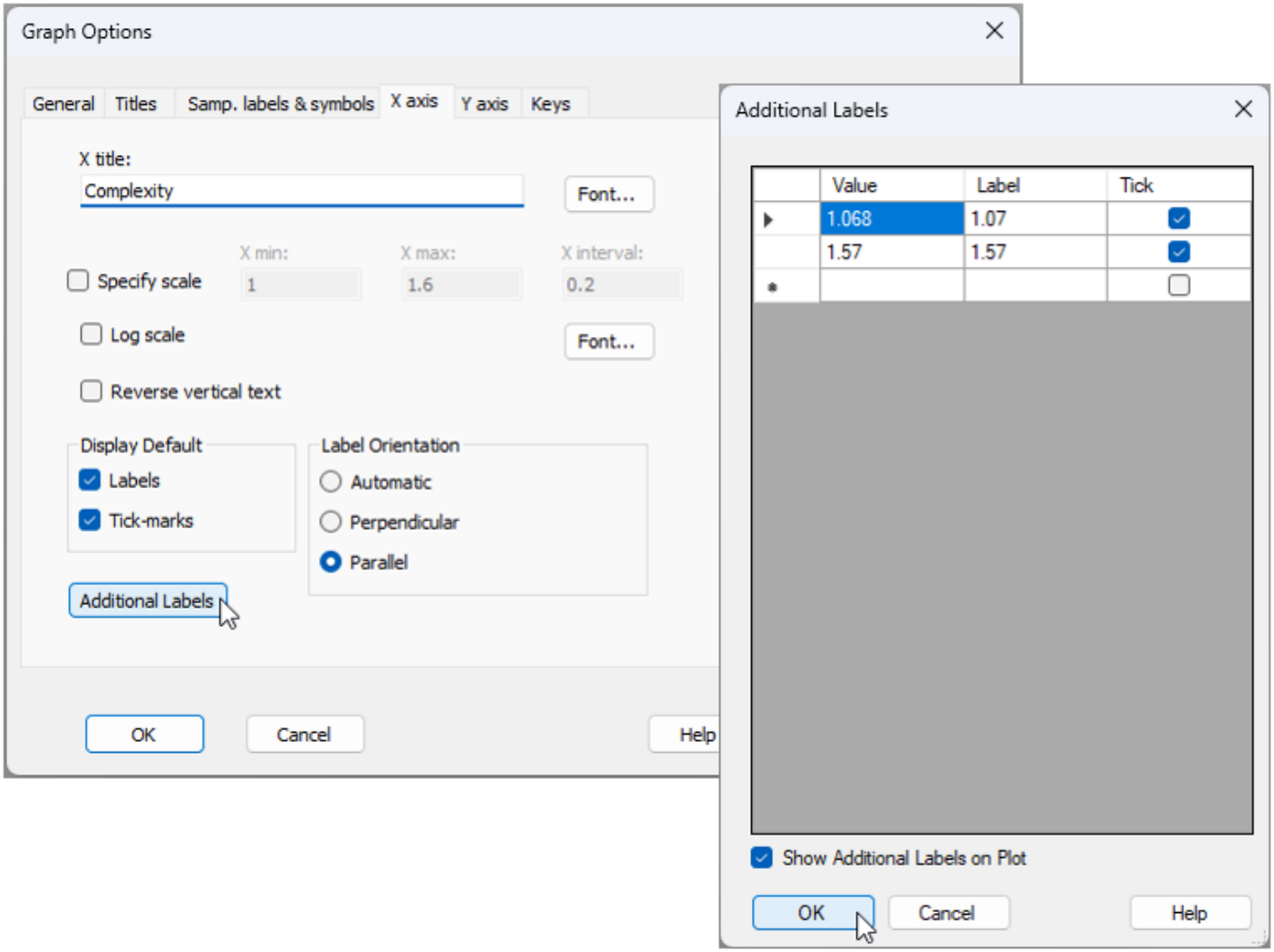
Consider the following scatter plot of species richness (S) *vs.* structural complexity of the substratum (measured using a chain-and-tape method), obtained from a visual survey of intertidal macrofauna and algae at a site inside the marine reserve at Long Bay, on the north shore of Auckland, New Zealand.¶ Biotic data ('Long_Bay_intertidal_biota') and environmental data ('Long_Bay_intertidal_env.pri') from this study are found in the 'Examples_P8' > 'Long_Bay_intertidal' folder.



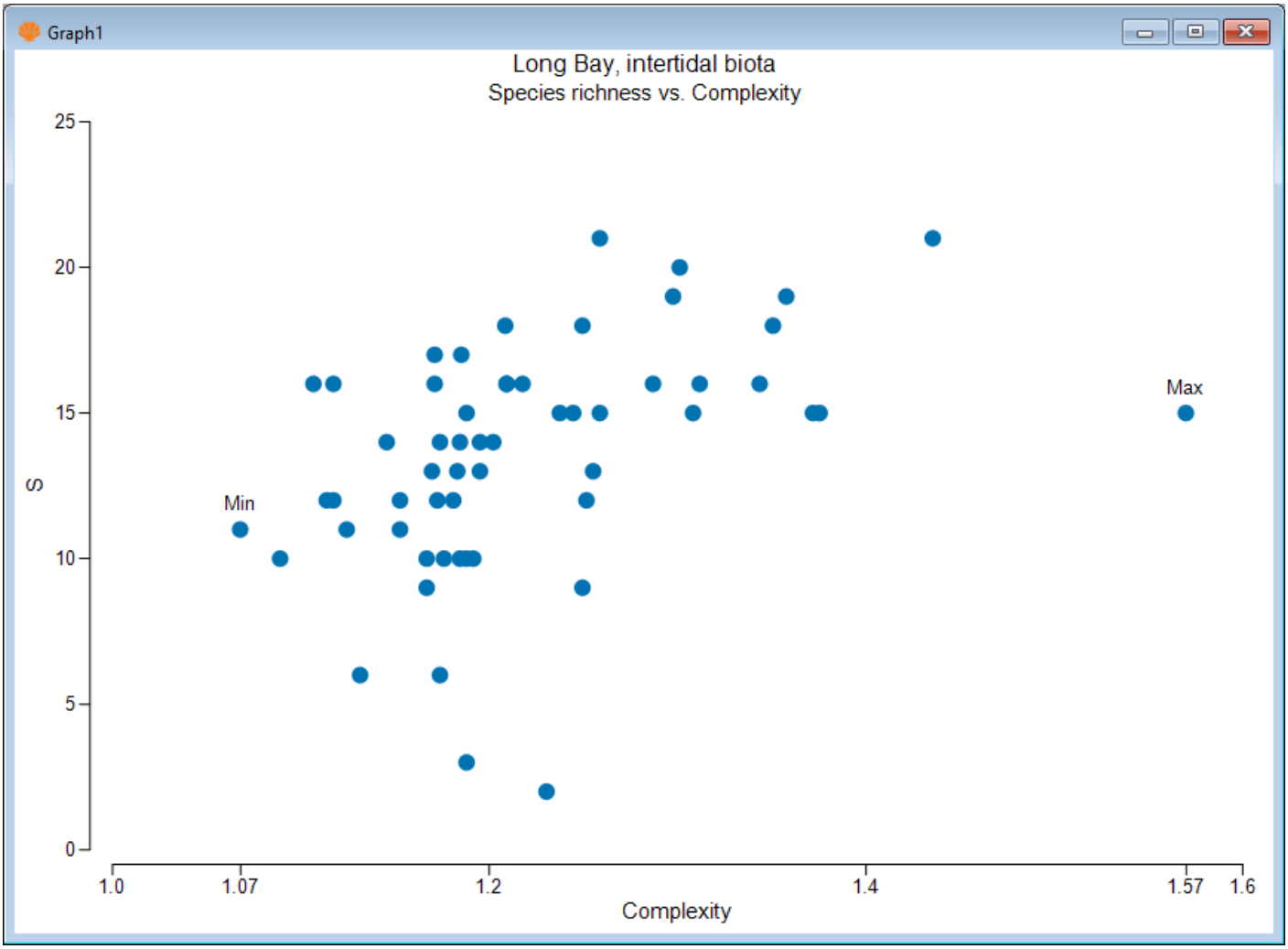
In the above scatter plot, the points that correspond to the minimum ('Min') and maximum ('Max') values obtained for the variable of 'Complexity' (on the x-axis), have been labeled individual (using a factor that is empty for all other samples). We might decide it would be useful to annotate the graphic to provide the minimum and maximum values on the x-axis for those two values directly as well. We can do this by clicking on the x-axis (or by clicking **Graph** > **General...** and clicking on

the 'X-axis' tab).

In this dialog, we can see an 'Additional Labels' button (Additional Labels), which is **new in P8**. We can also see the new option to change the **label orientation** so that it is either '$\bullet$Perpendicular' or '$\bullet$Parallel'. For the present example, let's choose '$\bullet$Parallel' and then go ahead and click the 'Additional Labels' button where we can add both the specific **values** and the desired **labels** for the maximum and minimum complexity, also including a tick-mark for each of them, like so:



The resulting graphic looks like this:



There are clearly many uses for this new feature. It is a great way to provide clarity regarding any specific values (or sets of values) that you may wish to highlight along particular axes in customised graphics.

Change the axis label orientation

Note that another new feature in PRIMER 8 is the possiblity to change the *orientation* of the axis labels. See section 3.2 for an example.

¶*These data were collected in 2015 as part of a course in quantitative marine ecology at Massey University, Albany, New Zealand. Taxa that were recorded in the survey as 'dead' (e.g., oyster shells, barnacle tests, etc.) were not included in the tally of richness analysed here.*

15.5 Split data sheet by factor/indicator

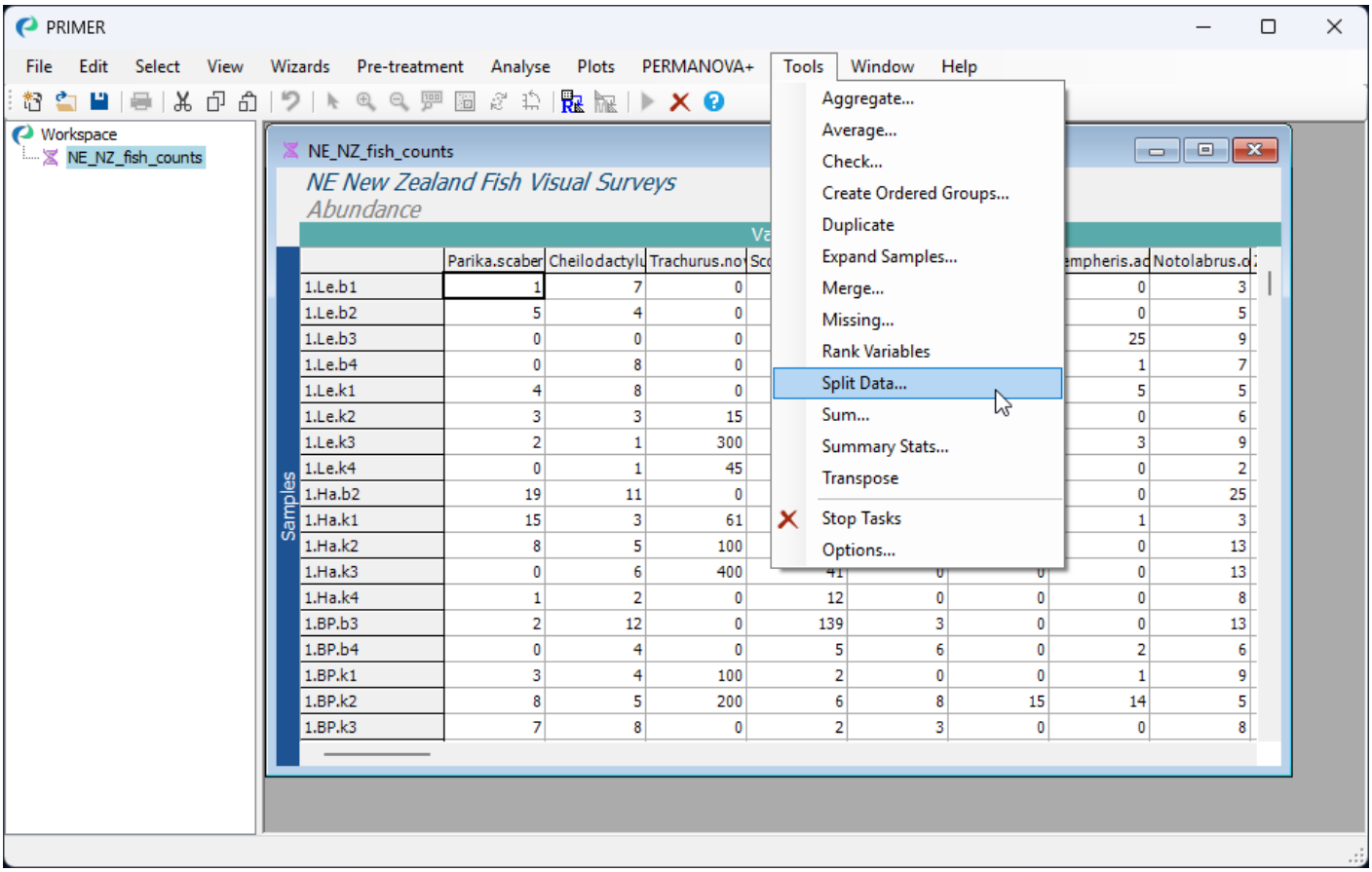
In PRIMER 8 there is a new tool, accessed by clicking **Tools** > **Split Data...**, which allows you to split a data sheet into several separate data sheets, corresponding to:

- separate groups of *samples*, based on a *factor*; or
- separate groups of *variables*, based on an *indicator*.

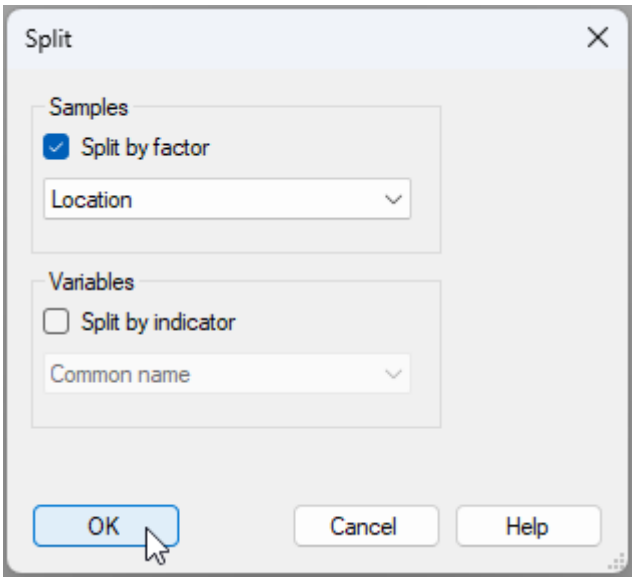
This is a simple tool, but a helpful one. Having this tool saves one from having to select each group individually, then duplicate them into a new sheet, one group at a time.

For example, consider the data set of fish surveys from New Zealand, discussed in section 10.4 above. (Data are located in the file 'NE_NZ_fish_counts.pri', found in the 'Example_P8' > 'NE_NZ_fish' folder.) Visual underwater surveys of fish assemblages were done at four different locations along the north-eastern coast of New Zealand: Berghan Point, Home Point, Leigh and Hahei. Suppose now we would like to examine trends in assemblages through time, but do this separately for each of these four different locations.

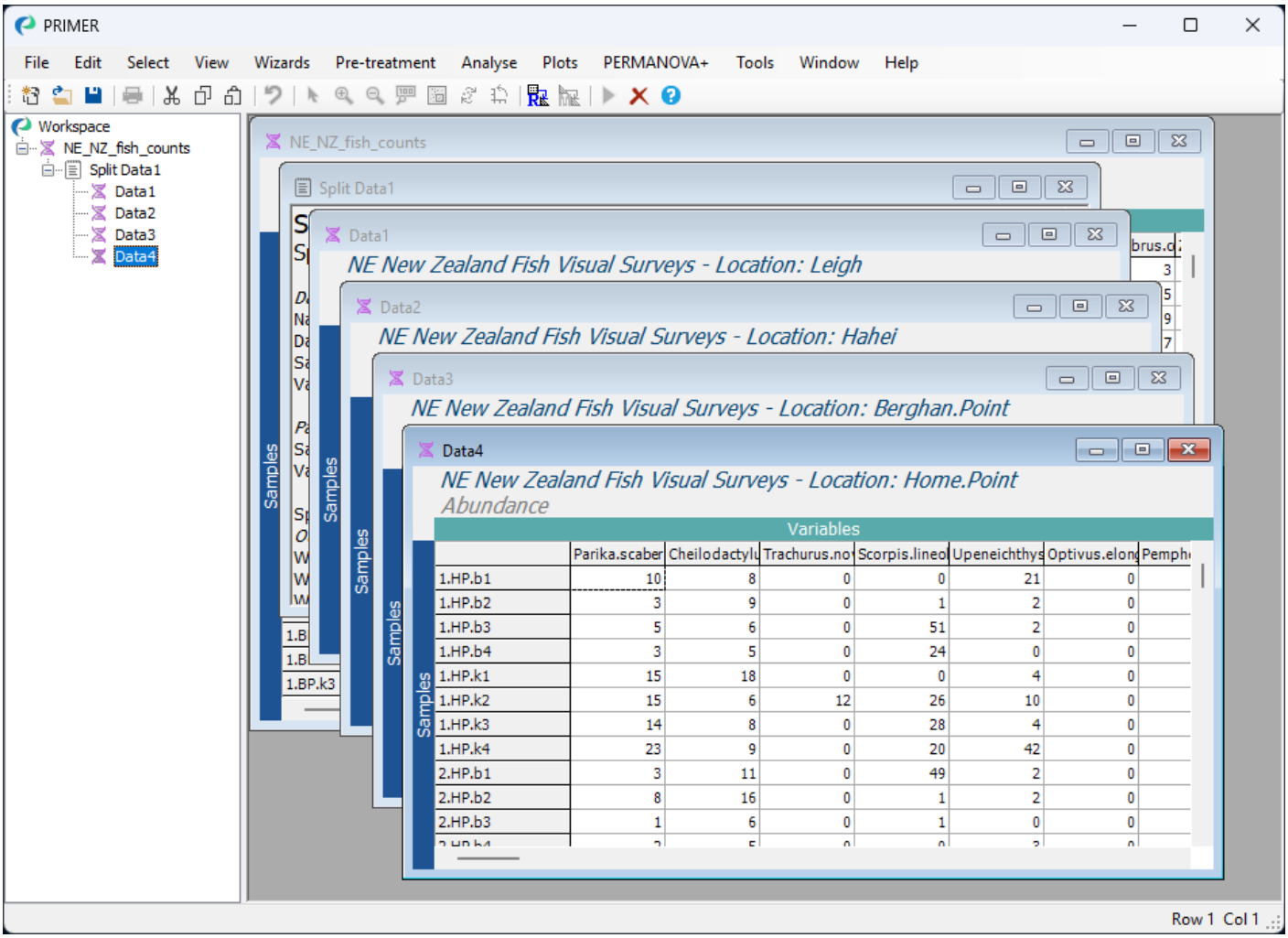
From the fish dataset, click **Tools** > **Split Data...**, like so:



We can then choose to split the data based on the factor of 'Location', as shown in the 'Split' dialog window below, then click '**OK**'.



In our Explorer tree, we then see four new data sheets, one for each of the four locations:



Note that the sheets produced by this operation each have a new *title* (see also **Edit** > **Properties**) that reflects:

- the factor on which the split was made; and

- the relevant corresponding group for each resulting data sheet.

For example, the title for 'Data4' is given as 'New Zealand Fish Visual Surveys - Location: Home.Point', *viz.:*

Of course, you can also change the name of the data sheets themselves if you wish. For example, you can change the name of the sheet called 'Data4' to 'Home.Point' by clicking **File** > **Rename Data**.

15.6 Line plots for samples

There is a new facility in PRIMER 8 to create **Line plots** in two different ways:

- with one line for every *variable* (across all samples); or
- with one line for every *sample* (across all variables).

This is a considerable improvement on the line plot dialog in PRIMER 7, which only could be implemented to draw lines for variables (Fig. 15.6).

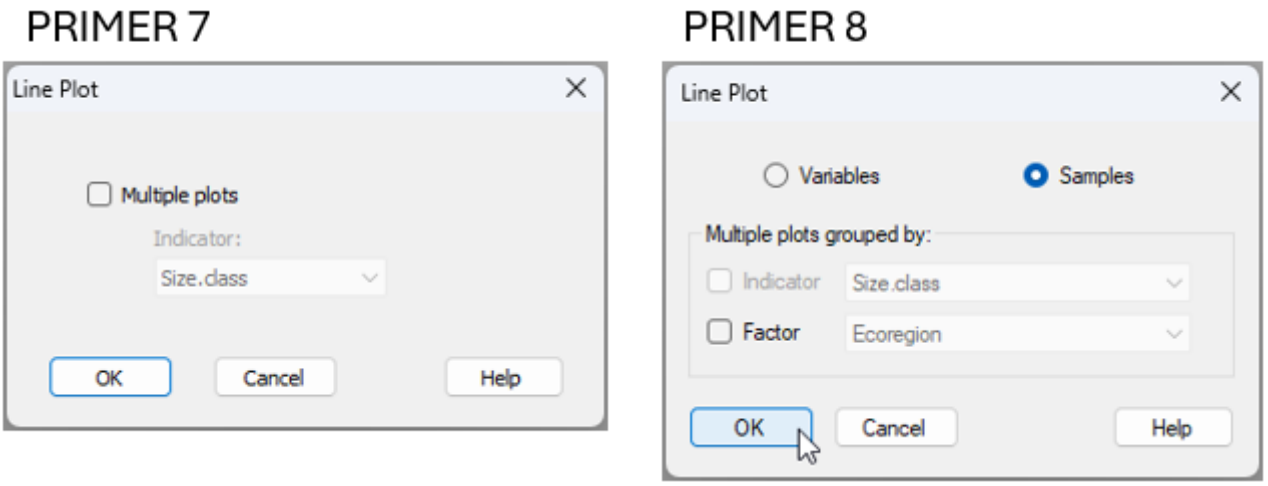
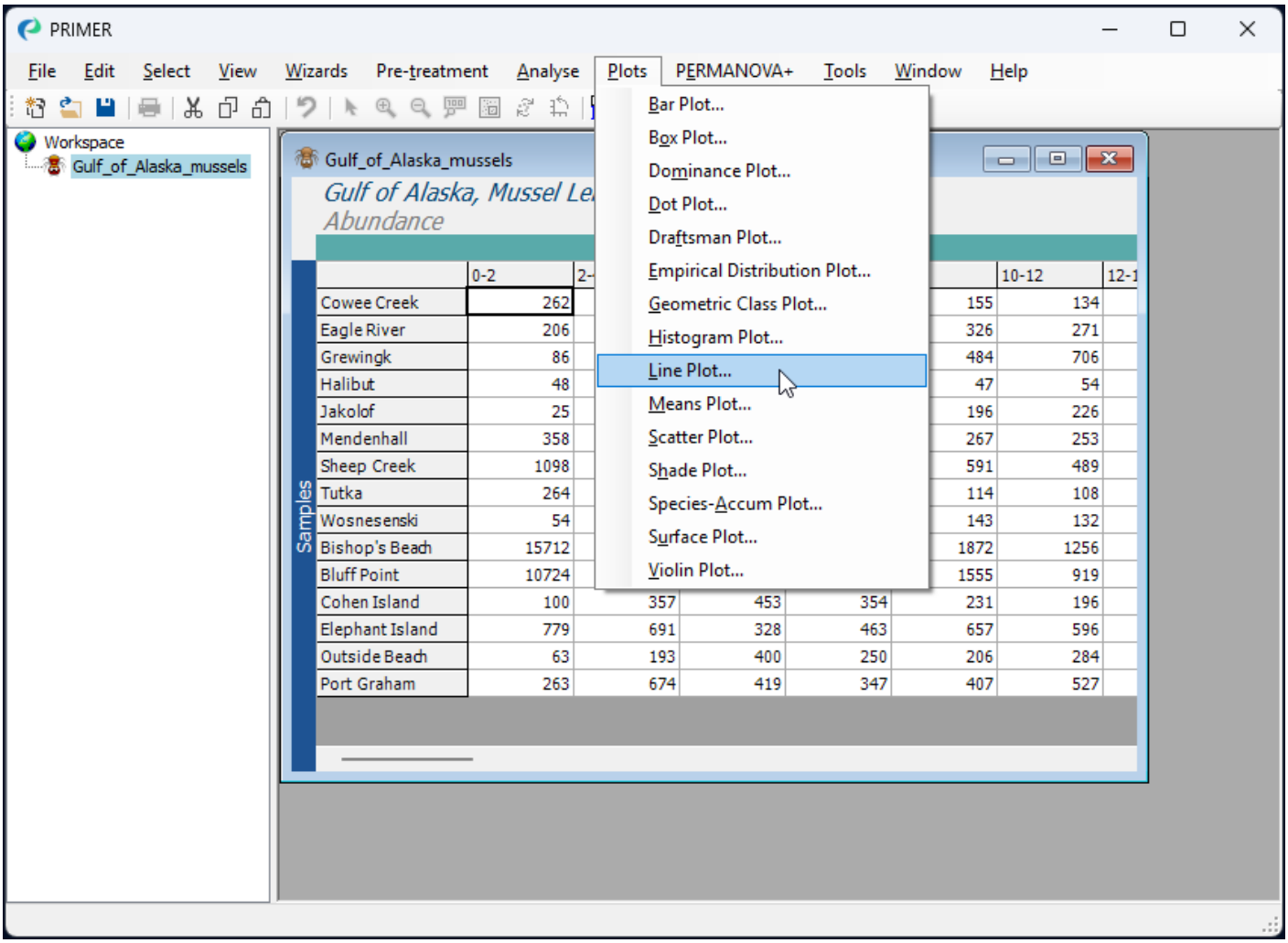


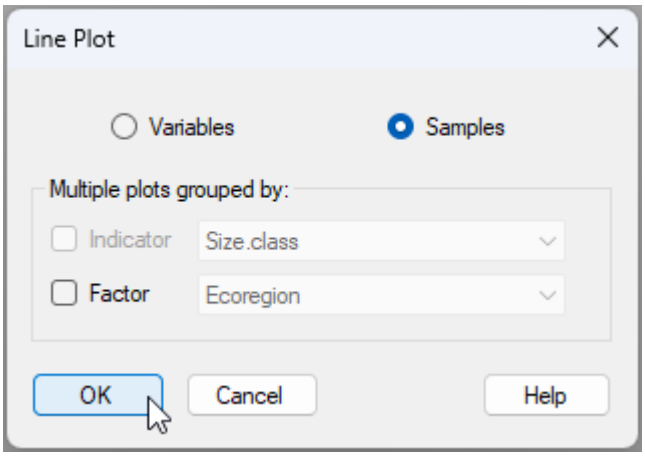
Fig. 15.6 Comparison of the Plots > Line Plot... dialog window in PRIMER 7 (at left) vs PRIMER 8 (at right).

We saw this tool earlier in section 13.3, in the analysis of mussel size-classes at different sites in the Gulf of Alaska (the data file is called 'Gulf_of_Alaska_mussels.pri', located in the folder 'Examples_P8' > 'Gulf_of_Alaska_mussels').

From the mussel data sheet, after standardising the original raw count data to cumulative percentages (the variables are the sizes classes here), resulting in a sheet called 'Data1', we can choose to create a line plot with a line for each sample (the sites here) by clicking **Plots** > **Line Plot...**, like so:

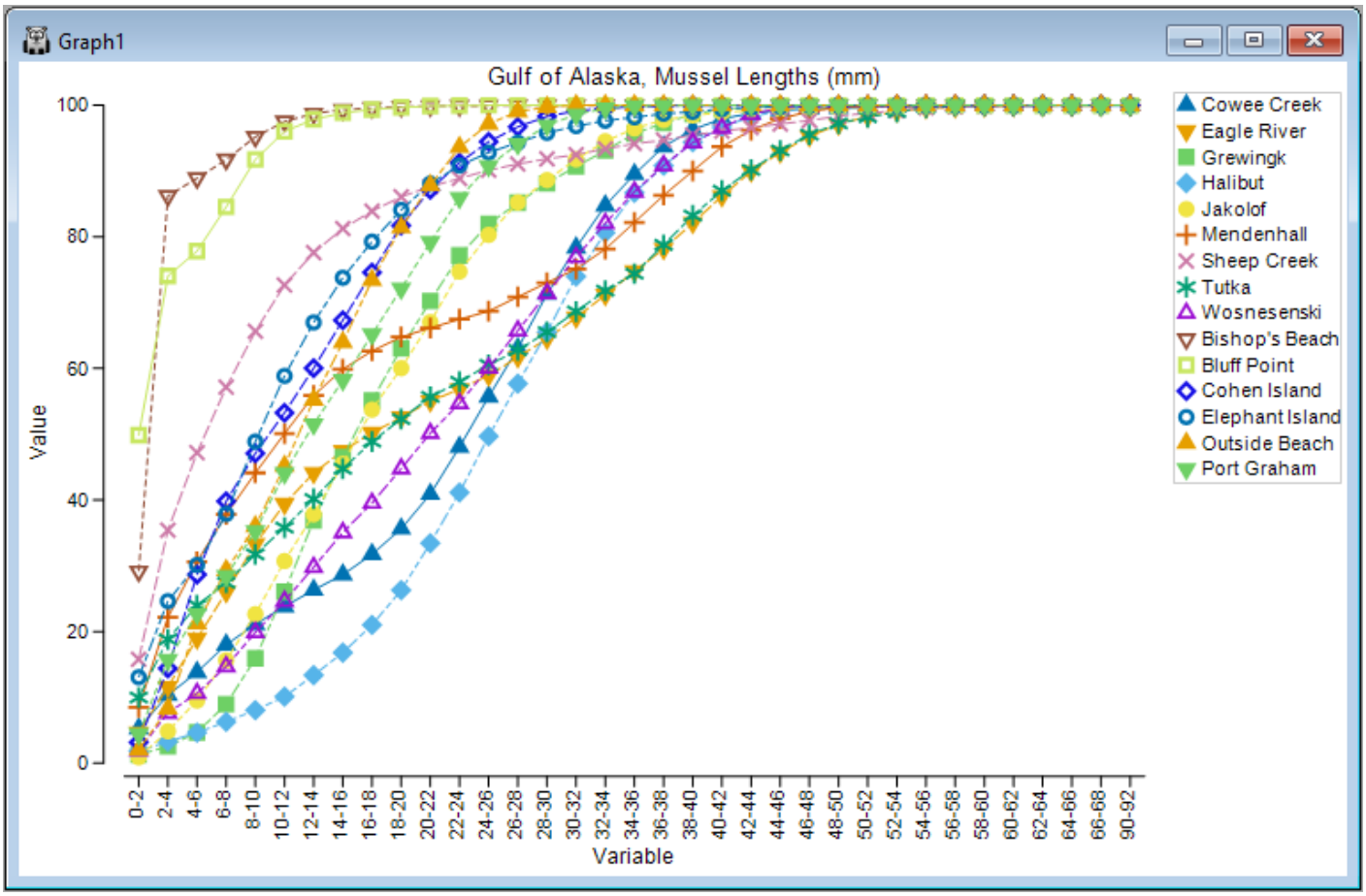


In the 'Line Plot'dialog window, we simply choose to view lines for the ($\bullet$ Samples), and then click '**OK**'.



Note: change the above dialog if the wording changes Note in the above dialog that we also have the option to output *multiple line plots* corresponding to groupsof samples identified *via* a factor (or groups of variables identified by an indicator, if we are drawing lines for variables).

The resulting line plot for this example is shown below:


Graph1
Gulf of Alaska, Mussel Lengths (mm)
Value
Variable
0
20
40
60
80
100
0-2
2-4
4-6
6-8
8-10
10-12
12-14
14-16
16-18
18-20
20-22
22-24
24-26
26-28
28-30
30-32
32-34
34-36
36-38
38-40
40-42
42-44
44-46
46-48
48-50
50-52
52-54
54-56
56-58
58-60
60-62
62-64
64-66
66-68
90-92
Cowee Creek
Eagle River
Grewingk
Halibut
Jakolof
Mendenhall
Sheep Creek
Tutka
Wosnesenski
Bishop's Beach
Bluff Point
Cohen Island
Elephant Island
Outside Beach
Port Graham

15.7 Output group-level stats from dispersion (or variability) weighting

In PRIMER 8, you can now output more detailed statistical information from either dispersion-weighting or variability weighting to a worksheet. You can see this additional option in the comparison of the 'Dispersion Weighting' dialog in PRIMER 7 (Fig. 15.7).

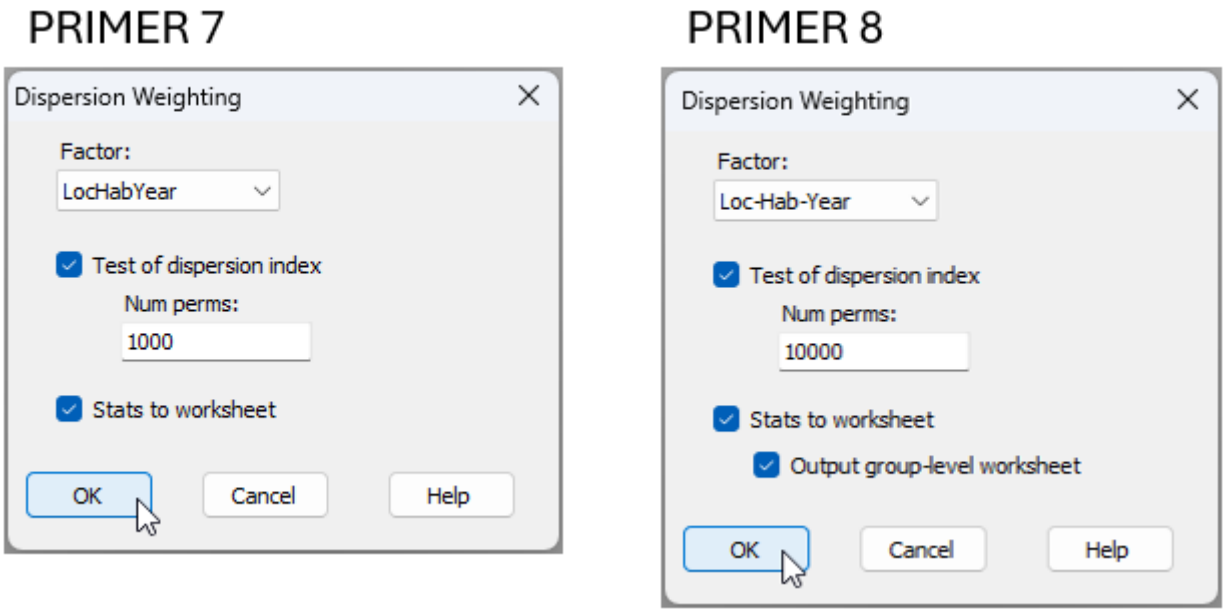


Fig. 15.7 Comparison of the Pre-treatment > Dispersion Weighting... dialog window in PRIMER 7 (at left) vs PRIMER 8 (at right).

In essence, whenever we calculate the average *index of dispersion* ($\bar{D}$) for the dispersion weighting pre-treatment, it is drawn from individual indices of dispersion (variance-to-mean ratios, D_i) for each of several ($i = 1, \ldots, g$) groups (identified by a factor). It would be helpful, in some situations, to be able to see those original individual indices of dispersion across all of the groups for each variable. The extra tickbox ('$\checkmark$Output group-level worksheet') available in the PRIMER 8 dispersion weighting dialog allows you to produce and examine this underlying statistical information in a worksheet directly.

For example, we can re-visit the data we saw in section 10.4 above, located in the file 'NE_NZ_fish_counts.pri', found in the 'Example_P8' > 'NE_NZ_fish' folder. We can begin by following steps 1, 2 and 3 from section 10.4 on these data; i.e., get the data into PRIMER 8, then:

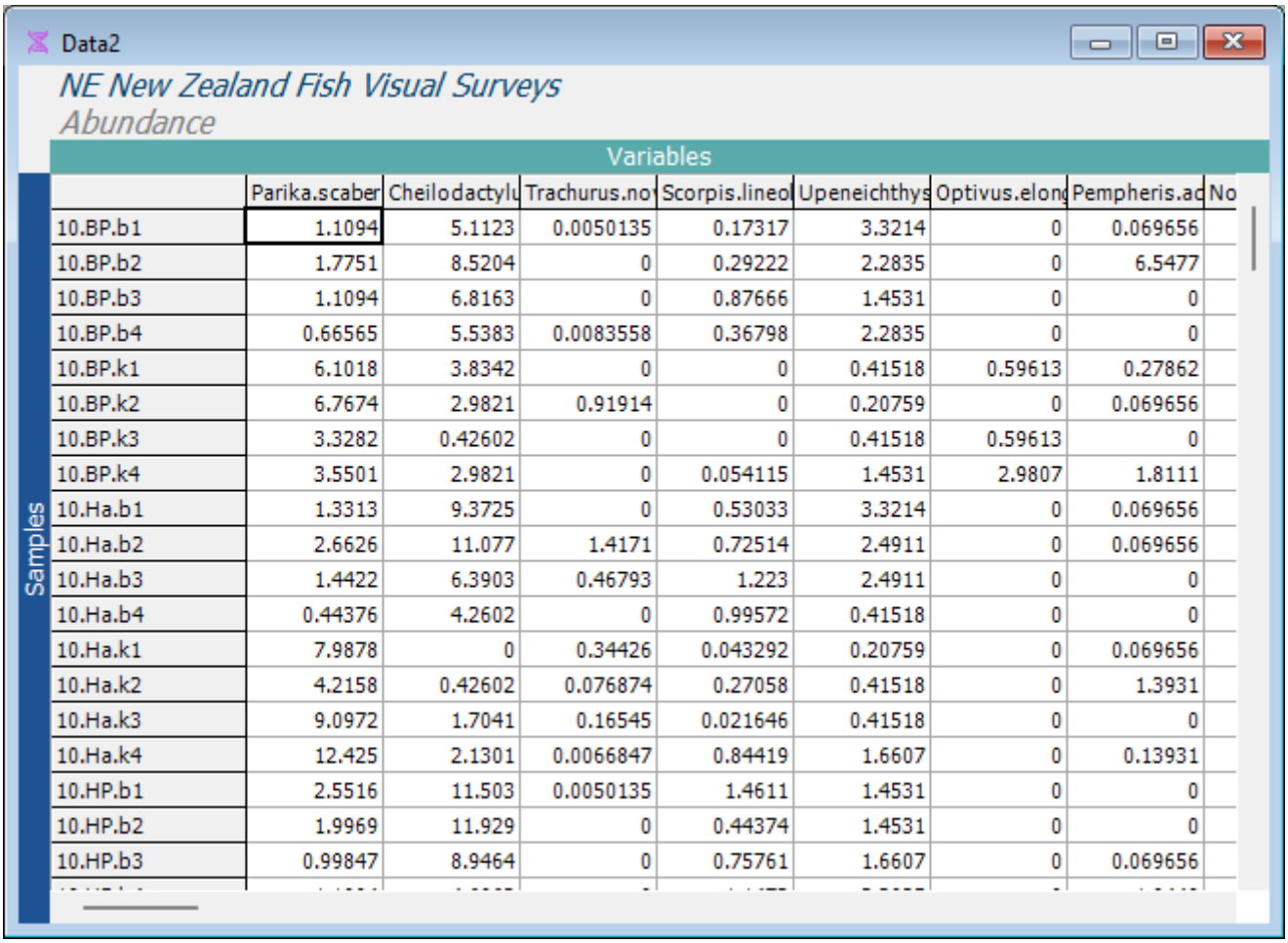
- Obtain a subset of the data based on the factor of 'Year', from 2010-2015 inclusive (**Select** > **Samples...**), with the resulting subset being called 'Data1'.
- Create a combined factor consisting of all combinations of the factors: 'Loc', 'Hab' and 'Year' (**Edit** > **Factors...** > **Combine...**)

Next, we can apply the dispersion weighting pre-treatment on the basis of this combined factor by clicking **Pre-treatment** > **Dispersion Weighting...** and in the resulting dialog, choose Factor: Loc-Hab-Year. This time, however, we shall also choose to get ($\checkmark$Stats to worksheet) & ($\checkmark$Output group-level worksheet), then click '**OK**'.

This will produce the following three worksheets (assuming 'Data1' contains the subset-selected data):

- 'Data2' contains the dispersion-weighted data.
- 'Data3' contains the statistics associated with the *average* index of dispersion ($\bar{D}$) for each variable.
- 'Data4' contains the *individual* indices of dispersion (D_i) for each group (columns) and for each variable (rows).

These three data sheets are shown below for the present example.



Data2

NE New Zealand Fish Visual Surveys

Abundance

Samples	Variables							
	Parika.scaber	Cheilodactylu	Trachurus.no	Scorpis.lineol	Upeneichthys	Optivus.elon	Pempheris.ac	No
10.BP.b1	1.1094	5.1123	0.0050135	0.17317	3.3214	0	0.069656	
10.BP.b2	1.7751	8.5204	0	0.29222	2.2835	0	6.5477	
10.BP.b3	1.1094	6.8163	0	0.87666	1.4531	0	0	
10.BP.b4	0.66565	5.5383	0.0083558	0.36798	2.2835	0	0	
10.BP.k1	6.1018	3.8342	0	0	0.41518	0.59613	0.27862	
10.BP.k2	6.7674	2.9821	0.91914	0	0.20759	0	0.069656	
10.BP.k3	3.3282	0.42602	0	0	0.41518	0.59613	0	
10.BP.k4	3.5501	2.9821	0	0.054115	1.4531	2.9807	1.8111	
10.Ha.b1	1.3313	9.3725	0	0.53033	3.3214	0	0.069656	
10.Ha.b2	2.6626	11.077	1.4171	0.72514	2.4911	0	0.069656	
10.Ha.b3	1.4422	6.3903	0.46793	1.223	2.4911	0	0	
10.Ha.b4	0.44376	4.2602	0	0.99572	0.41518	0	0	
10.Ha.k1	7.9878	0	0.34426	0.043292	0.20759	0	0.069656	
10.Ha.k2	4.2158	0.42602	0.076874	0.27058	0.41518	0	1.3931	
10.Ha.k3	9.0972	1.7041	0.16545	0.021646	0.41518	0	0	
10.Ha.k4	12.425	2.1301	0.0066847	0.84419	1.6607	0	0.13931	
10.HP.b1	2.5516	11.503	0.0050135	1.4611	1.4531	0	0	
10.HP.b2	1.9969	11.929	0	0.44374	1.4531	0	0	
10.HP.b3	0.99847	8.9464	0	0.75761	1.6607	0	0.069656	

Data3

Index of Dispersion (D) Coefficients - Loc-Hab-Year

Other

Samples

Variables

	D	Sig%	P-Value	Divisor
Parika.scaber	9.0138	1.5835E-181	1.5835E-183	9.0138
Cheilodactylus.spectabilis	2.3473	1.5682E-15	1.5682E-17	2.3473
Trachurus.novaezelandiae	598.39	0	0	598.39
Scorpis.lineolatus	92.396	0	0	92.396
Upeneichthys.lineatus	4.8172	2.3859E-70	2.3859E-72	4.8172
Optivus.elongatus	10.065	0	0	10.065
Pempheris.adspersus	14.356	2.1293E-284	2.1293E-286	14.356
Notolabrus.celidotus	5.9996	2.3007E-100	2.3007E-102	5.9996
Zeus.faber	0.93333	100	1	1
Chromis.dispilus	151.16	0	0	151.16
Chironemus.marmoratus	1.8168	0	0	1.8168
Chrysophrys.auratus	8.2934	1.0567E-161	1.0567E-163	8.2934
Aldrichetta.forsteri	1	100	1	1
Caesioperca.lepidoptera	17.514	0	0	17.514
Nemadactylus.douglasii	1.4472	0.03	0.0003	1.4472
Myliobatis.tenuicaudatus	1	49.83	0.4983	1
Odax.pullus	3.1285	0	0	3.1285
Scorpis.violaceus	32.499	0	0	32.499
Coris.sandageri	4.7521	0	0	4.7521
Girella.tricuspidata	33.29	0	0	33.29
Pseudolabrus.luculentus	1.6	1.65	0.0165	1.6
Pseudolabrus.miles	1.6952	0	0	1.6952

Data4

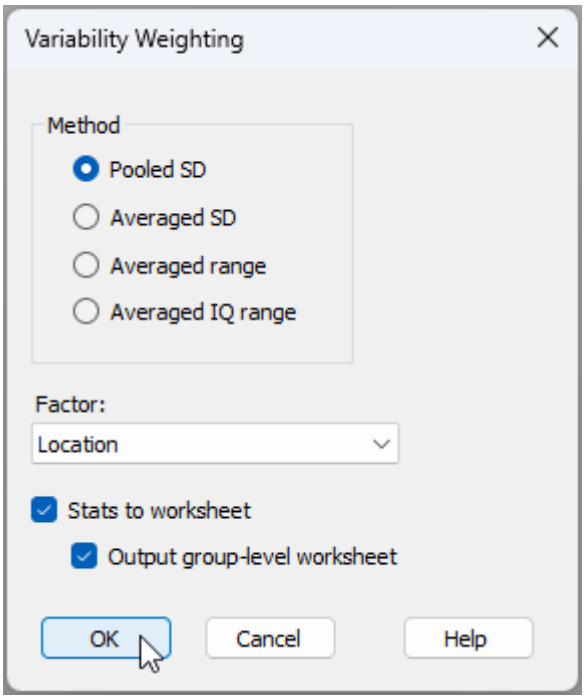
Indices of Dispersion (D) per Loc-Hab-Year Group

Other

Variables	Samples							
	BP-b-2010	BP-k-2010	Ha-b-2010	Ha-k-2010	HP-b-2010	HP-k-2010	Le-b-2010	Le
Parika.scaber	1.619	5.6105	5.1006	12.246	2.9778	4.381	1.3111	
Cheilodactylus.spec	0.84699	2	2.79	2.2667	2.8008	1.619	4.9333	
Trachurus.novaezel	3	550	566.68	85.869	3	1186.9	Missing!	
Scorpis.lineolatus	20.768	5	9.8141	45.979	19.189	9.6173	187.49	
Upeneichthys.lineat	1.2074	2.4444	3.3968	3.1538	0.43434	6.2	9.0364	
Optivus.elongatus	Missing!	16.857	Missing!	Missing!	Missing!	5	Missing!	
Pempheris.adspersu	92.361	19.473	0.66667	15.812	13.5	1	19.961	
Notolabrus.celidotus	0.27451	0.52941	18.304	0.18095	4.7566	2.6	4.0241	
Zeus.faber	Missing!	Missing!	Missing!	Missing!	Missing!	Missing!	Missing!	
Chromis.dispilus	159.41	219.8	106.92	346.27	482.19	476.59	7.2105	
Chironemus.marmo	1.6923	Missing!	4.6889	1	0.066667	1	2	
Chrysophrys.auratus	2.7576	9.7778	8	15.72	8.2444	0.48718	3.4737	
Aldrichetta.forsteri	Missing!	Missing!	Missing!	Missing!	Missing!	Missing!	Missing!	
Caesioperca.lepidop	Missing!	Missing!	Missing!	Missing!	Missing!	Missing!	Missing!	
Nemadactylus.doug	2	0.73333	Missing!	Missing!	2.5556	0.52381	1	
Myliobatis.tenuicau	Missing!	Missing!	1	1	1	2	Missing!	
Odax.pullus	21	1.0606	5	0.61404	Missing!	3.1905	Missing!	
Scorpis.violaceus	48.524	12.588	81.872	8	52.692	56.723	Missing!	
Coris.sandageri	7.4035	0.33333	2	4	0.68627	6.381	Missing!	

Note in the above sheet ('Data4', containing individual indices, D_i) that there are many missing values. This is simply a consequence of there being zero fish of that species in that particular group of samples; hence, the mean and the variance will both be equal to zero, so no value of D_i can be calculated for that particular cell (or group). Average $\bar{D}$ values for each species are naturally calculated only from those cells (groups of samples) where at least some individuals of that species were recorded (i.e., the non-missing entries).

Finally, note that this '$\checkmark$Output group-level worksheet' option is also available for **Pre-treatment** > **Variability Weighting...**, *viz.:*


Variability Weighting
Method
Pooled SD
Averaged SD
Averaged range
Averaged IQ range
Factor:
Location
Stats to worksheet
Output group-level worksheet
OK
Cancel
Help

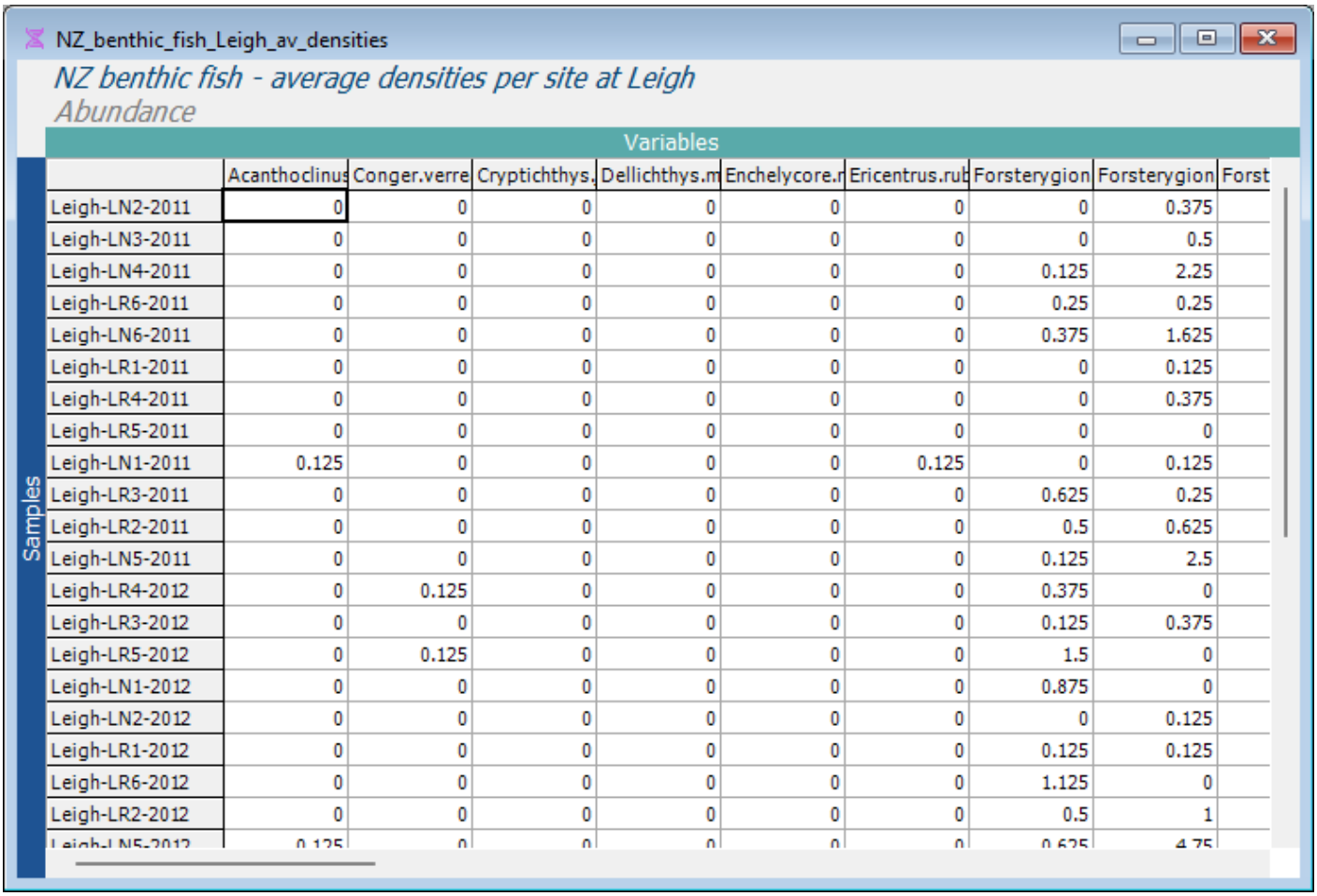
15.8 Output diagnostic plots from CAP

In PRIMER 8, it is now possible to output diagnostic plots from the CAP routine (i.e., canonical analysis of principal coordinates, an analysis obtained from a resemblance matrix by clicking **PERMANOVA+** > **CAP**). You simply tick the new option to '$\checkmark$Do diagnostic plots' in the CAP dialog window.

For example, consider a study of the assemblages of small benthic fishes living in rocky subtidal reef habitats in northeastern New Zealand, examined by Smith & Anderson (2016) . Data from this study are located in the file called 'NZ_benthic_fish.pri', in the 'Examples_P8' > 'NZ_benthic_fish' folder. Surveys were done of the benthic fish fauna, along with fine-scale habitat features in kelp forests and rocky reefs along the north-eastern coast of New Zealand. There were sites at a range of locations in and around several marine reserves (including Leigh, Tawharanui, Hahei and the Poor Knights Islands), and data were obtained over a period of 3 years (2011-2013). At each site, divers surveyed n = 8 transects, measuring 1 m $\times$ 5 m. Each transect was made up of five contiguous 1 m $\times$ 1 m quadrats. For each quadrat, divers first visually searched and recorded counts for all benthic fishes, then recorded the presence/absence of a set of pre-defined habitat features (see the file named 'NZ_benthic_fish_habitat.pri', located in the same folder).

Here, we shall consider only the potential effects of the marine reserve at Leigh on these fish communities in a fully balanced design. Note that the factor of 'Reserve' has two levels: 0 = outside the reserve, and 1 = inside the reserve. Due to the sparsity of the count data at small spatial scales, we shall analyse data summed to the transect level, then consider densities of each species (averages per transect) at the spatial scale of sites for the ensuing CAP analysis. Mean densities (per $5m^2$) of each fish species per site are located in the file named ' NZ_benthic_fish_Leigh_av_densities.pri'.

1. Open the file ('NZ_benthic_fish_Leigh_av_densities.pri') in PRIMER 8. It will look like this:


NZ_benthic_fish_Leigh_av_densities

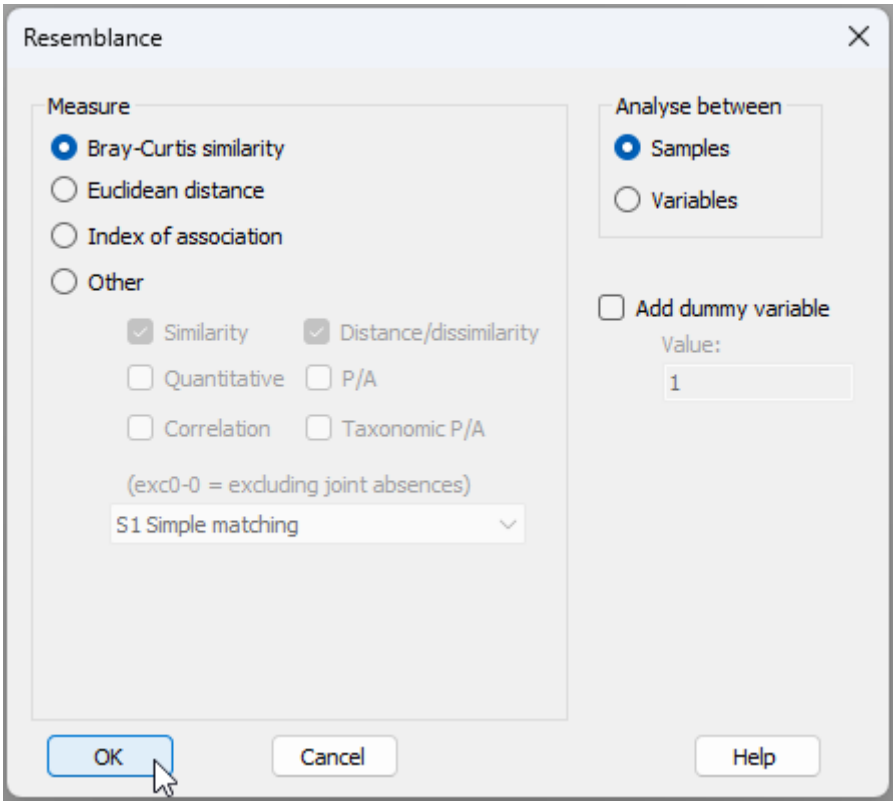
NZ benthic fish - average densities per site at Leigh

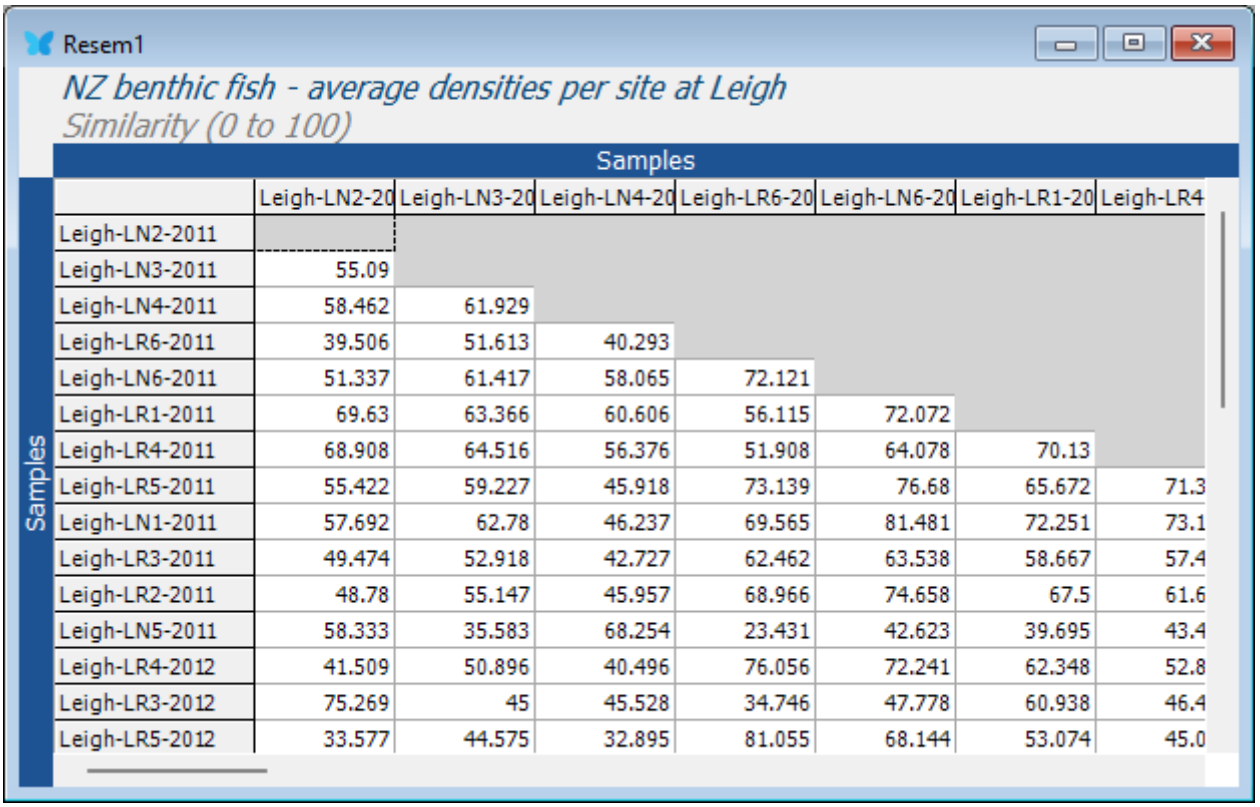
Abundance

Variables

Samples	Acanthoclinus	Conger.verre	Cryptichthys.	Dellichthys.m	Enchelycore.r	Ericentrus.rub	Forsterygion	Forsterygion	Forst
Leigh-LN2-2011	0	0	0	0	0	0	0	0.375	
Leigh-LN3-2011	0	0	0	0	0	0	0	0.5	
Leigh-LN4-2011	0	0	0	0	0	0	0.125	2.25	
Leigh-LR6-2011	0	0	0	0	0	0	0.25	0.25	
Leigh-LN6-2011	0	0	0	0	0	0	0.375	1.625	
Leigh-LR1-2011	0	0	0	0	0	0	0	0.125	
Leigh-LR4-2011	0	0	0	0	0	0	0	0.375	
Leigh-LR5-2011	0	0	0	0	0	0	0	0	
Leigh-LN1-2011	0.125	0	0	0	0	0.125	0	0.125	
Leigh-LR3-2011	0	0	0	0	0	0	0.625	0.25	
Leigh-LR2-2011	0	0	0	0	0	0	0.5	0.625	
Leigh-LN5-2011	0	0	0	0	0	0	0.125	2.5	
Leigh-LR4-2012	0	0.125	0	0	0	0	0.375	0	
Leigh-LR3-2012	0	0	0	0	0	0	0.125	0.375	
Leigh-LR5-2012	0	0.125	0	0	0	0	1.5	0	
Leigh-LN1-2012	0	0	0	0	0	0	0.875	0	
Leigh-LN2-2012	0	0	0	0	0	0	0	0.125	
Leigh-LR1-2012	0	0	0	0	0	0	0.125	0.125	
Leigh-LR6-2012	0	0	0	0	0	0	1.125	0	
Leigh-LR2-2012	0	0	0	0	0	0	0.5	1	
Leigh-LN5-2012	0.125	0	0	0	0	0	0.625	4.75	

2. From 'NZ_benthic_fish_Leigh_av_densities', click **Analyse** > **Resemblance...** and choose to calculate the Bray-Curtis similarity among samples. This will produce a resemblance matrix called 'Resem1'.




Resem1

NZ benthic fish - average densities per site at Leigh

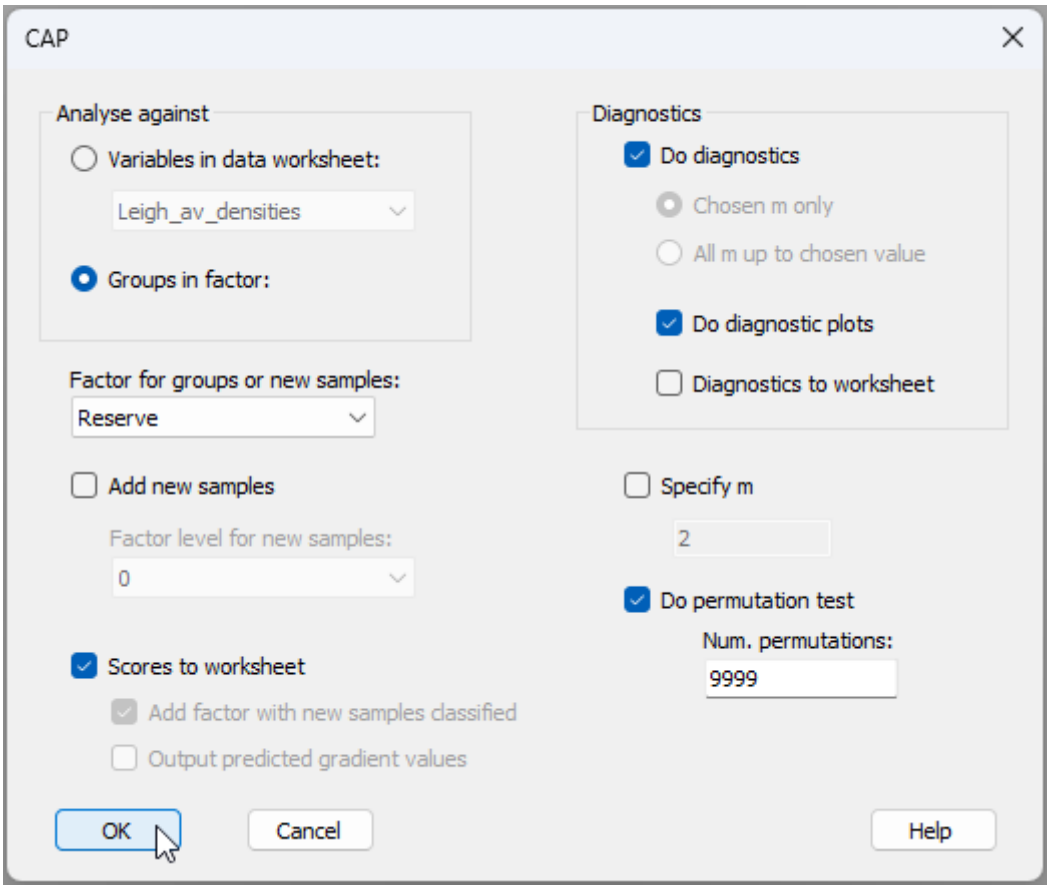
Similarity (0 to 100)

Samples

Samples	Leigh-LN2-20	Leigh-LN3-20	Leigh-LN4-20	Leigh-LR6-20	Leigh-LN6-20	Leigh-LR1-20	Leigh-LR4
Leigh-LN2-2011							
Leigh-LN3-2011	55.09						
Leigh-LN4-2011	58.462	61.929					
Leigh-LR6-2011	39.506	51.613	40.293				
Leigh-LN6-2011	51.337	61.417	58.065	72.121			
Leigh-LR1-2011	69.63	63.366	60.606	56.115	72.072		
Leigh-LR4-2011	68.908	64.516	56.376	51.908	64.078	70.13	
Leigh-LR5-2011	55.422	59.227	45.918	73.139	76.68	65.672	71.3
Leigh-LN1-2011	57.692	62.78	46.237	69.565	81.481	72.251	73.1
Leigh-LR3-2011	49.474	52.918	42.727	62.462	63.538	58.667	57.4
Leigh-LR2-2011	48.78	55.147	45.957	68.966	74.658	67.5	61.6
Leigh-LN5-2011	58.333	35.583	68.254	23.431	42.623	39.695	43.4
Leigh-LR4-2012	41.509	50.896	40.496	76.056	72.241	62.348	52.8
Leigh-LR3-2012	75.269	45	45.528	34.746	47.778	60.938	46.4
Leigh-LR5-2012	33.577	44.575	32.895	81.055	68.144	53.074	45.0

3. From the 'Resem1' matrix, run the CAP routine aiming to distinguish small benthic fish communities inside *vs* outside the marine reserve, by clicking **PERMANOVA+** > **CAP** and choose:
   (Analyse against $\bullet$Groups in factor > Factor for groups or new samples: 'Reserve') &
   ($\checkmark$Scores to worksheet) &
   (Diagnostics > $\checkmark$Do diagnostics > $\checkmark$Do diagnostic plots) &
   ($\checkmark$Do permutation test > Num. permutations: 9999)

then click **OK**, as shown below:



The option to *do diagnostic plots* is new to PRIMER 8.

The analysis will run and produce the following:

- 'CAP1' (CAP1), containing all of the essential results in a rich text format (*.rtf), and
- 'MultiPlot1' (MultiPlot1), containing the CAP plot itself (as 'Graph1') along with plots of the *diagnostics for choosing m*, where m = the number of principal coordinate axes (PCOs) used to produce the canonical (CAP) axis (as 'Graph2' through 'Graph5').

From 'CAP1' and the accompanying CAP plot ('Graph1'), shown below, we can see that the canonical correlation associated with the 'Reserve' effect is reasonably strong (δ_1 = 0.644), and that this is statistically significant (δ_1^2 = 0.415, P = 0.001). This CAP model was achieved with m = 4 PCO axes, which achieved a leave-one-out allocation success of 77.78% under cross-validation.

CAP1

CANONICAL ANALYSIS

Correlations

Eigenvalue	Correlation	Corr.Sq.
1	0.644	0.4147

DIAGNOSTICS

m	prop.G	ssres	d_1^2	%correct
1	0.4148	0.99277	0.0626	61.111
2	0.5625	0.74153	0.3958	72.222
3	0.6935	0.66856	0.411	75
4	0.7783	0.65859	0.4147	77.778
5	0.8554	0.68304	0.4397	75
6	0.9145	0.74106	0.4428	72.222
7	0.9528	0.74104	0.4496	75
8	0.9902	0.84985	0.4507	69.444

Cross Validation

Leave-one-out Allocation of Observations to Groups
(for the choice of m: 4)

	classified			
Orig. group	0	1	Total	%correct
0	14	4	18	77.778
1	4	14	18	77.778

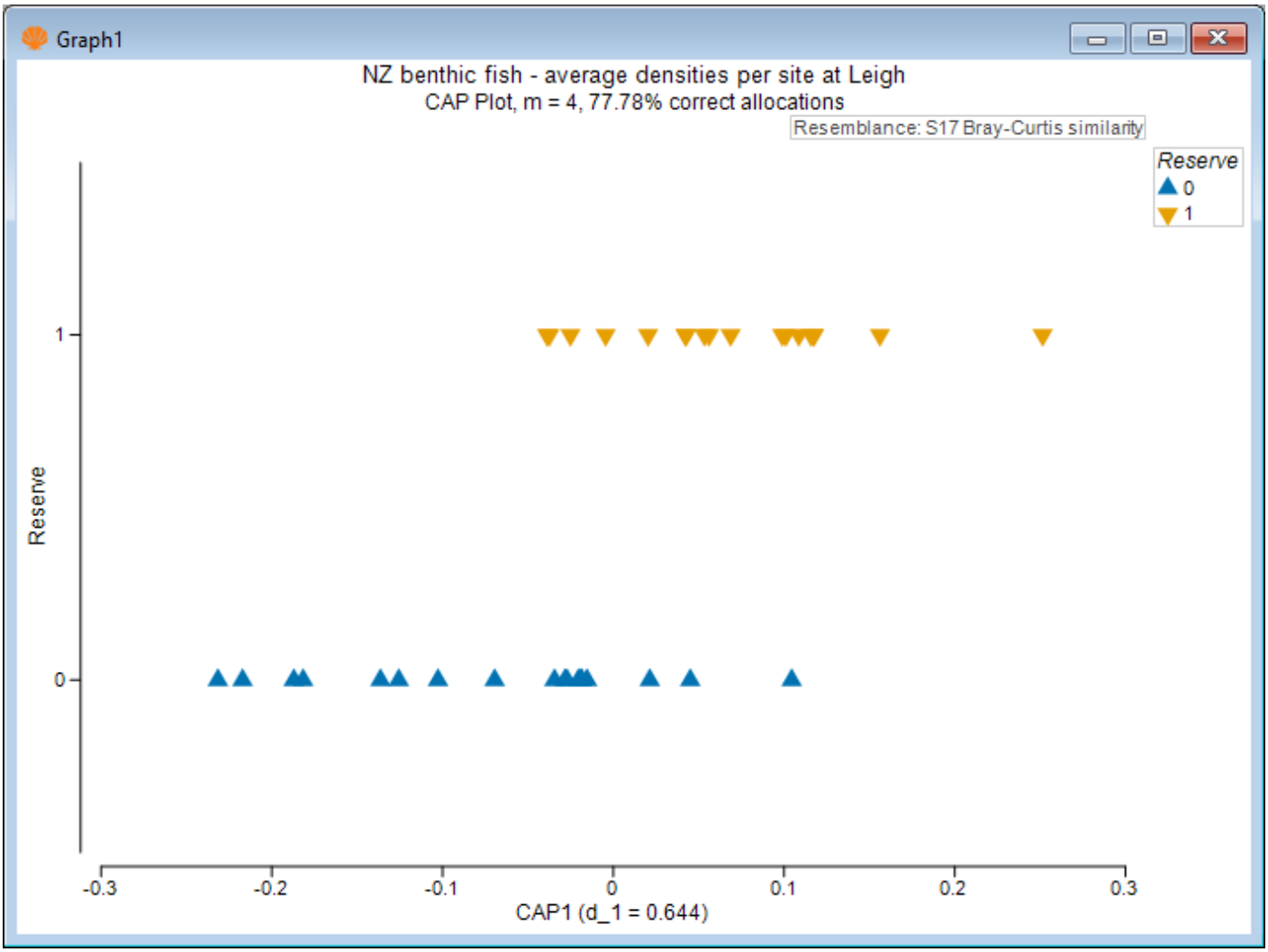
Total correct: 28/36 (77.778%)
Mis-classification error: 22.222%

Individual samples that were mis-classified

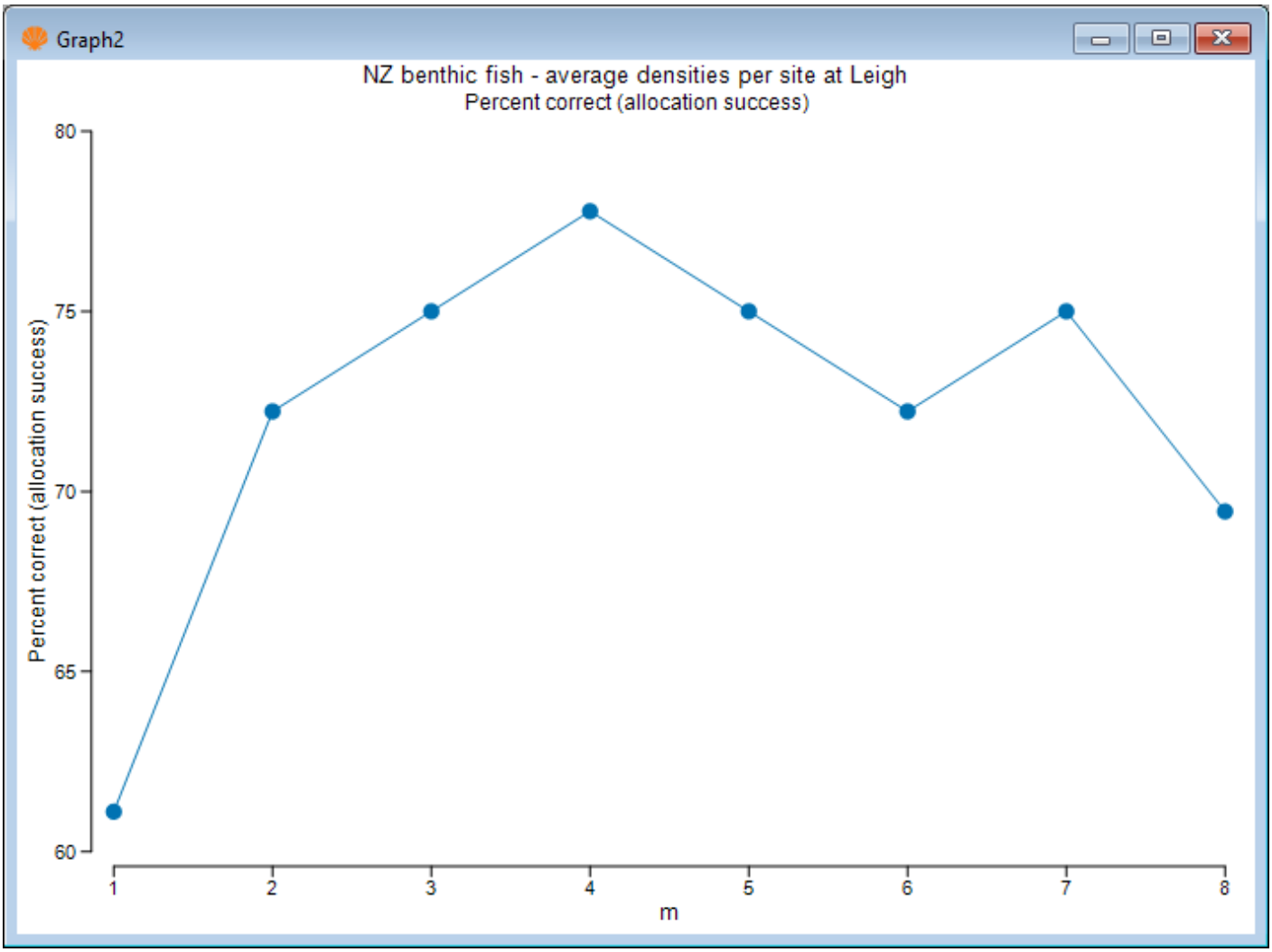
Sample	Orig.group	Class.group
Leigh-LN1-2011	0	1
Leigh-LN1-2012	0	1
Leigh-LN1-2013	0	1
Leigh-LN2-2013	0	1
Leigh-LR1-2011	1	0
Leigh-LR3-2012	1	0
Leigh-LR2-2012	1	0
Leigh-LR2-2013	1	0

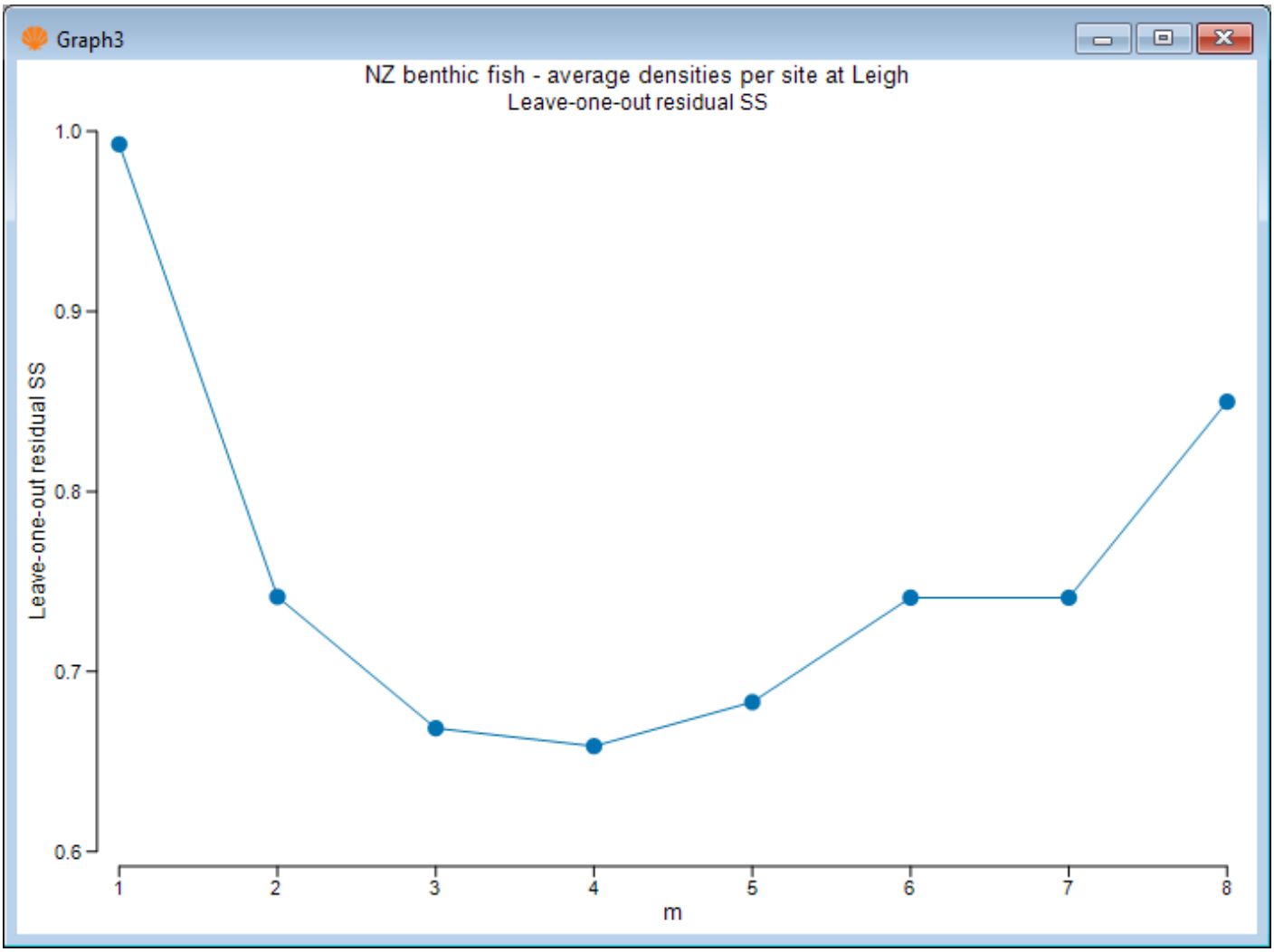
PERMUTATION TEST
trace statistic = (tr(Q_m'HQ_m))
first squared canonical correlation = (delta_1^2)

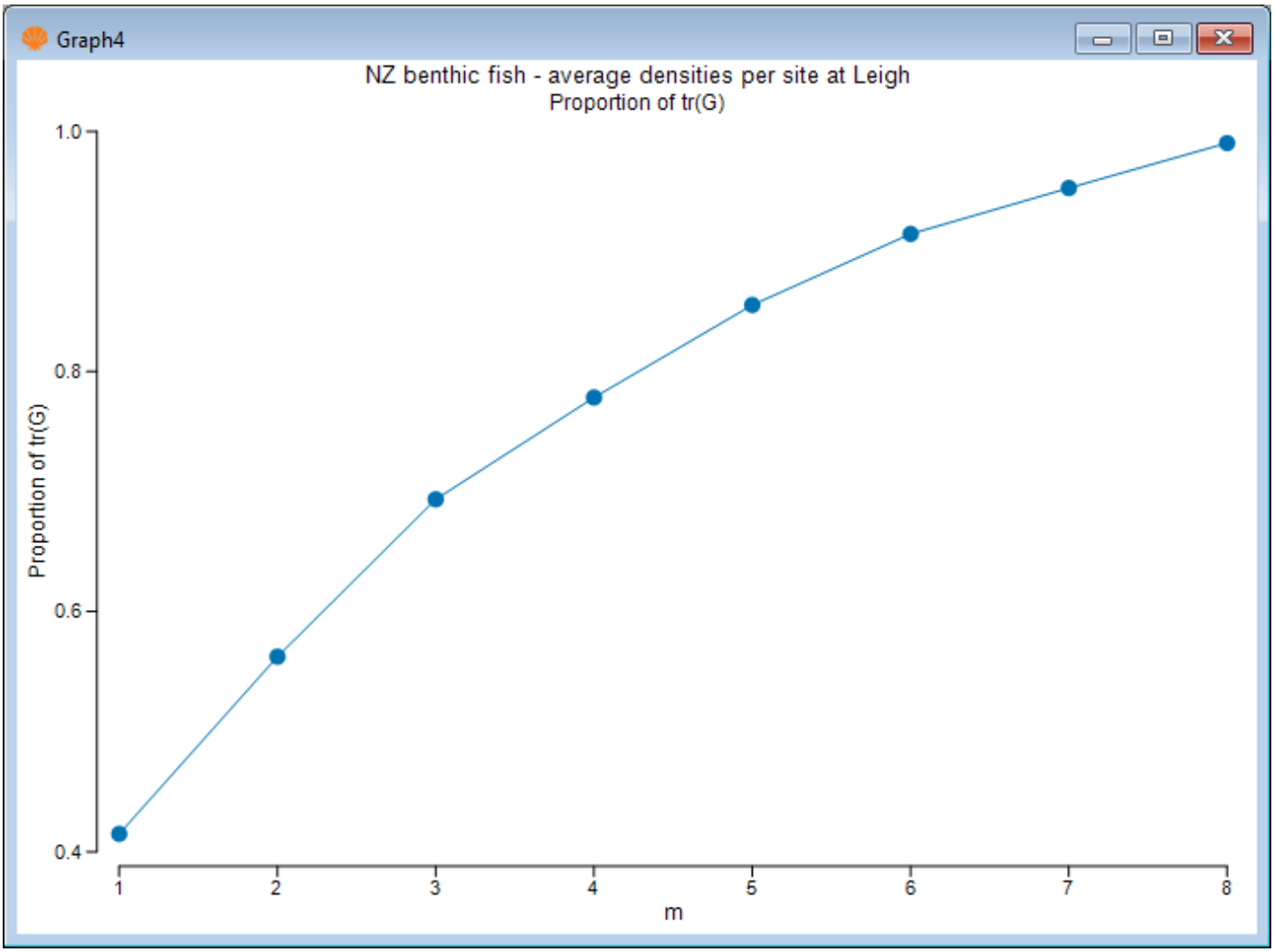
tr(Q_m'HQ_m): 0.41472 P: 0.001
delta_1^2: 0.41472 P: 0.001
No. of permutations used: 9999

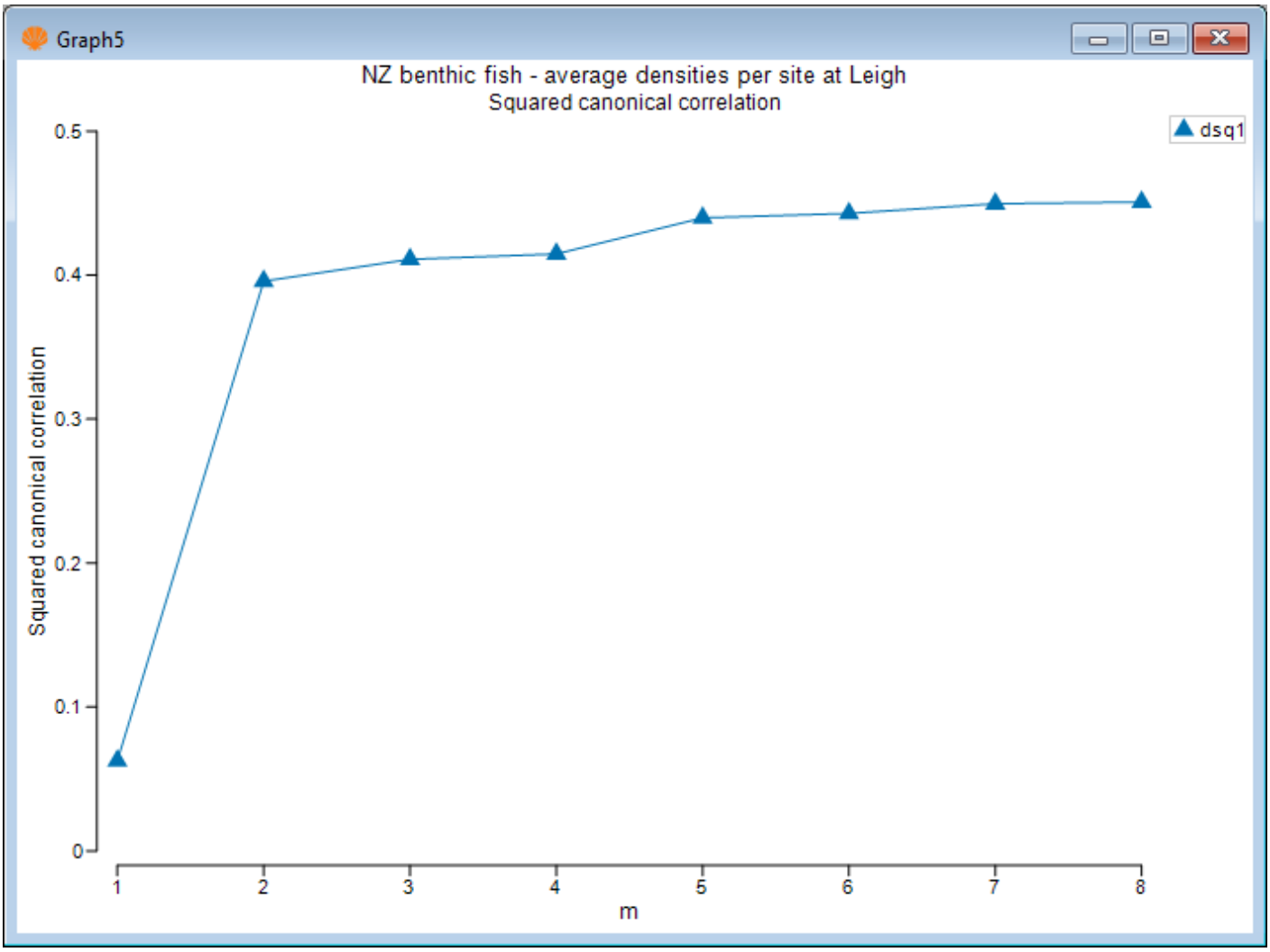


The four *diagnostic plots* are shown below.


Graph2
NZ benthic fish - average densities per site at Leigh
Percent correct (allocation success)
Percent correct (allocation success)
80
75
70
65
60
1
2
3
4
5
6
7
8
m


Graph3
NZ benthic fish - average densities per site at Leigh
Leave-one-out residual SS
Leave-one-out residual SS
1.0
0.9
0.8
0.7
0.6
1
2
3
4
5
6
7
8
m


Graph4
NZ benthic fish - average densities per site at Leigh
Proportion of tr(G)
1.0
0.8
0.6
0.4
Proportion of tr(G)
1
2
3
4
5
6
7
8
m



It is clear that the choice of m = 4 is a good one here, as this choice achieved the greatest leave-one-out allocation success ('Graph2') and the lowest leave-one-out residual sum-of-squares ('Graph3'). Although this can also (technically) be seen in the '*DIAGNOSTICS*' section of the 'CAP1' output above, it is certainly much clearer to see this information in diagnostic plots, whereby the wisdom of the choice made, overall, by reference to other values of m, can be readily assessed.

15.9 New diagnostics for PCA/PCO plots

Background

Consider a cloud of N points (sampling units) in a p-dimensional multivariate space. Principal components analysis (PCA) is an ordination method that will find an axis through that cloud of points so as to maximise the total variance of the points, when they are projected at right angles onto that axis. Having found such an axis, a second axis is then built in a similar way, subject to it being perpendicular to (independent of) the first axis, and so on. PCA does this in Euclidean space. Principal coordinate analysis (abbreviated as PCO or PCoA) also does this, but in the space of a chosen resemblance measure (Gower (1966)).

Hence, we may conceptually consider either PCA or PCO as a type of *projection* of the points into a smaller number of dimensions that can capture important major stuctures occurring across the data cloud as a whole. Specifically, if there is some redundancy of information (inter-correlation) among the p original variables, then we may anticipate that a subset of (say) two or three principal axes will capture a substantial portion of the total variation in the system. A configuration plot of the positions of the points along the first two (or three) principal axes can therefore be examined in an attempt to visualise patterns among the sampling units along these major axes of variation.

To assess the utility of a 2-d (or 3-d) PCA/PCO plot, one might simply examine the percentage of the total variation that is extracted by the first 2 (or 3) axes. If this is substantial (e.g., 70% or 80%), then we may consider the plot to be useful for interpretation. Another useful tool is a *scree plot*, which shows the percentage of variation explained by sequentially ordered principal axes.

This does not, however, provide us with a meaningful measure of how well the plot (of reduced dimension) represents the (high-dimensional) inter-point distances (or dissimilarities) in the original multivariate space.

A good way to do that would be to calculate the *stress* associated with the 2-d (or 3-d) configuration. The notion of stress is already quite familiar as a measure of the adequacy of an ordination plot produced using multi-dimensional scaling (MDS, see section 5.2 and section 5.8 in *Change in Marine Communities, 3rd edition*). Indeed, it is the specific goal of the MDS algorithm to create a configuration that minimises stress. Of course, neither PCA nor PCO are specifically designed to minimise stress, so a PCA or PCO configuration will always (necessarily) have higher stress than an MDS configuration for a given data set. Nevertheless, having a measure of stress for a PCA/PCO configuration can help us assess its utility for displaying inter-point relationships

faithfully. Here, we would consider that the usual rule-of-thumb of stress < 0.20 indicates a usefully interpretable configuration. Also, a plot of the configuration distances *vs* the original inter-point distances - a Shepard diagram - is clearly also a useful diagnostic tool here.

Scree Plot, Shepard diagram and Stress

In the PCO routine (accessed from a resemblance matrix by clicking on **PERMANOVA+** > **PCO...**), PRIMER 8 now offers you new options to output these diagnostics tools:

- a Scree plot;
- Shepard diagrams in 2D and/or 3D; and
- Stress calculated using either metric MDS or threshold metric MDS

Fig. 15.8 compares the dialog window for the PCO routine in PRIMER 7 versus PRIMER 8.

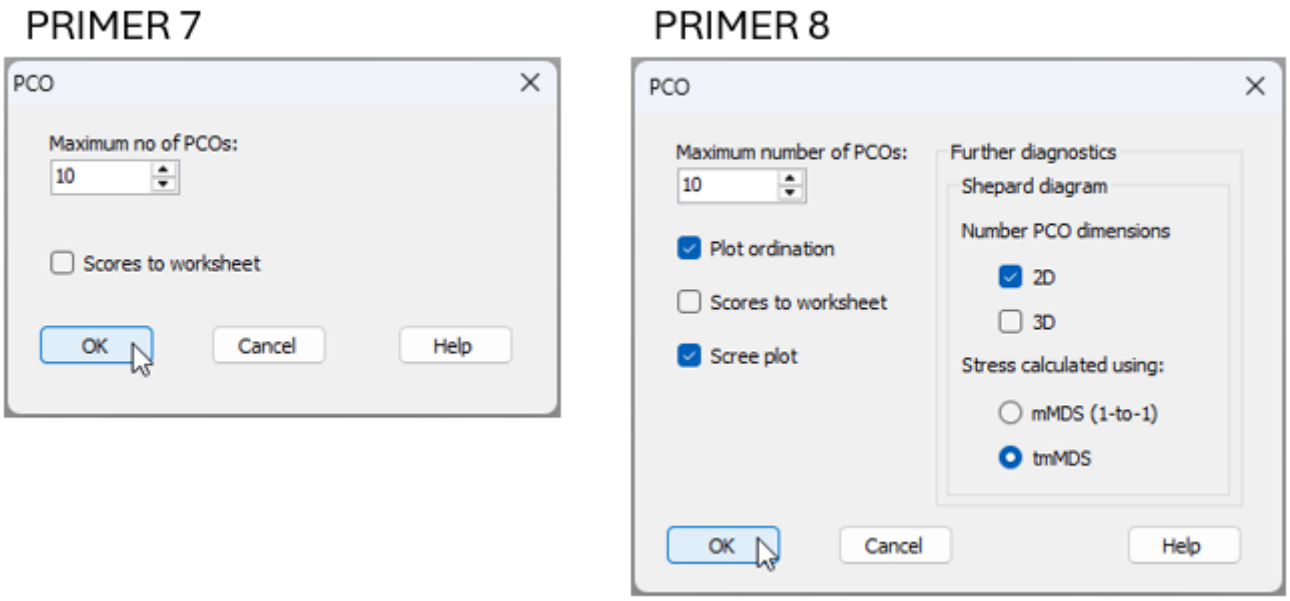


Fig. 15.8 Comparison of the PERMANOVA+ > PCO... dialog window in PRIMER 7 (at left) vs PRIMER 8 (at right).

These same new diagnostic output options are also available for the PCA routine in PRIMER 8 (obtained from a data matrix by clicking on **Analyse** > **PCA...**).

It is important to recognise that these additional diagnostics do not in any way change anything about the fundamental PCA or PCO analysis itself that is being done. The calculation of principal axes for PCA or PCO ordination still remains exactly the same as ever, and may best be thought of as a *projection*. What is new and different here is simply the opportunity also to view the stress and the Shepard diagrams which attend the resulting PCA or PCO configuration. These diagnostics also help put these methods into perspective vis-à-vis any MDS ordination solutions one might obtain for a given dataset.

Example: Messolongi lagoon diatoms

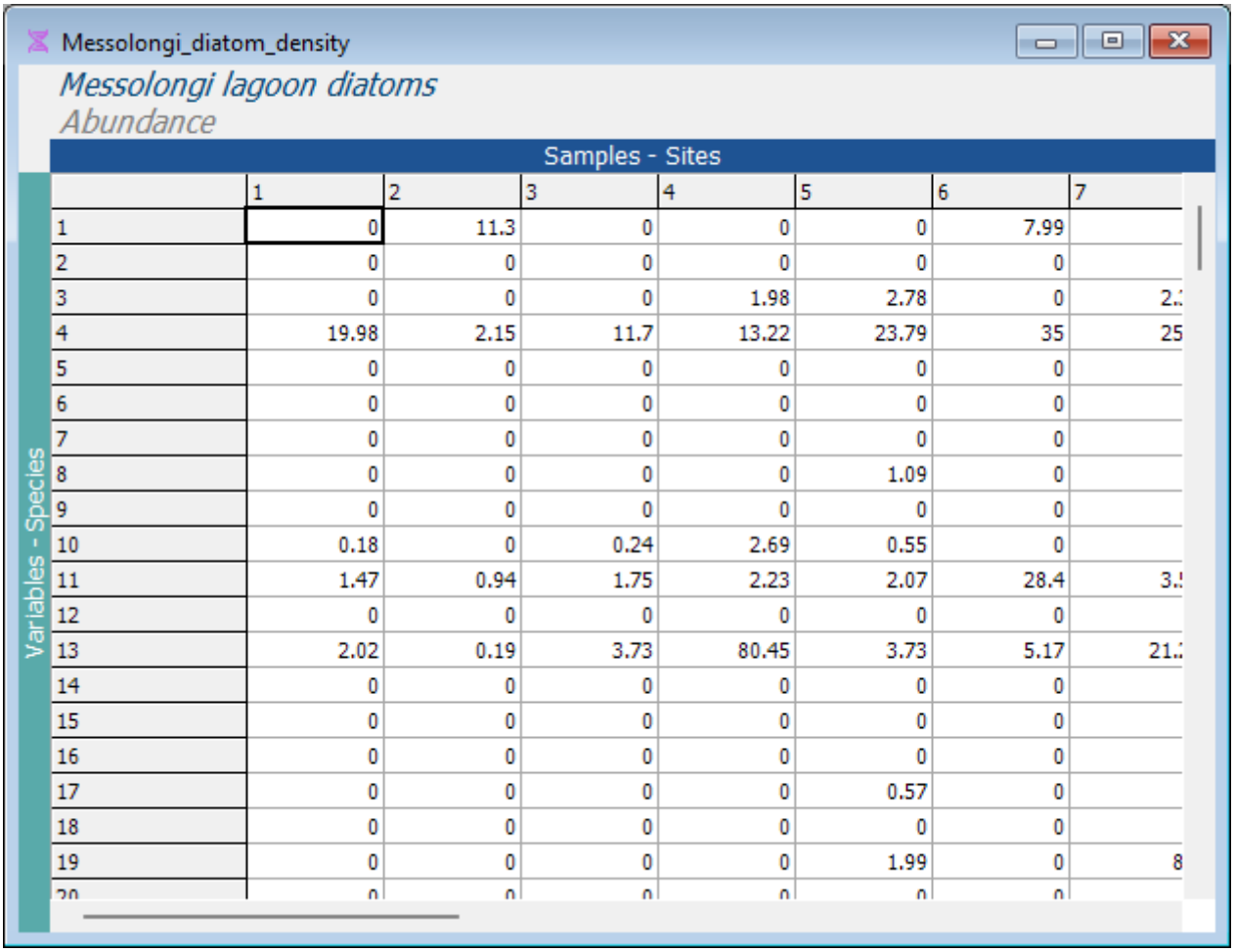
Let's consider a study of diatom assemblages (densities of 193 species) at 17 sites in the lagoons of Messolongi, Aitoliko and Kleissova in Eastern Central Greece (Danielidis (1991)). Data from this study are contained in the file called 'Messolongi_diatom_density.pri', located in the 'Examples_P8' > 'Messolongi_diatoms' folder.

In what follows, we shall create an ordination of these data on the basis of a Bray-Curtis resemblance matrix calculated from square-root transformed densities, using each of the following methods:

- Non-metric MDS (nMDS)
- Threshold-metric MDS (tmMDS)
- Metric MDS (mMDS); and
- Principal co-ordinates analysis (PCO).

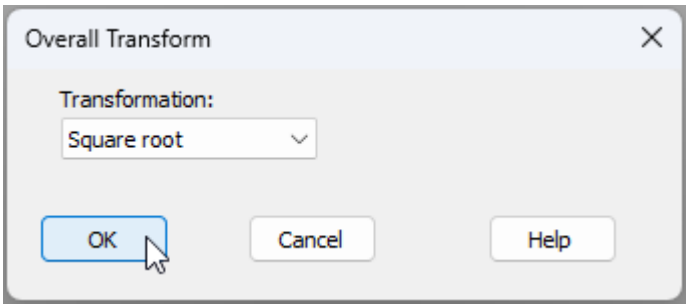
We will compare and contrast not only the ordination diagrams produced, but also the associated diagnostic Shepard diagrams and the values of stress. For PCO, we can produce Shepard diagrams and stress values in two different ways: (i) for comparison with tmMDS; or (ii) for comparison with mMDS.¶

1. Open up the data matrix (Messolongi_diatom_density.pri) in PRIMER 8.

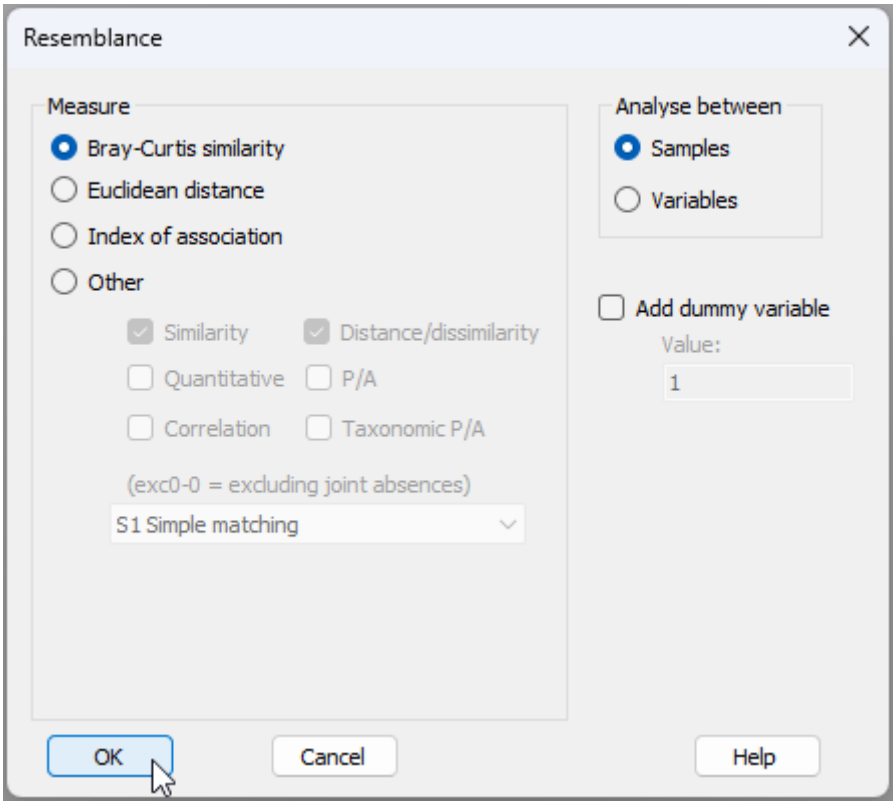


	1	2	3	4	5	6	7
1	0	11.3	0	0	0	7.99	
2	0	0	0	0	0	0	
3	0	0	0	1.98	2.78	0	2.
4	19.98	2.15	11.7	13.22	23.79	35	25
5	0	0	0	0	0	0	
6	0	0	0	0	0	0	
7	0	0	0	0	0	0	
8	0	0	0	0	1.09	0	
9	0	0	0	0	0	0	
10	0.18	0	0.24	2.69	0.55	0	
11	1.47	0.94	1.75	2.23	2.07	28.4	3.
12	0	0	0	0	0	0	
13	2.02	0.19	3.73	80.45	3.73	5.17	21.
14	0	0	0	0	0	0	
15	0	0	0	0	0	0	
16	0	0	0	0	0	0	
17	0	0	0	0	0.57	0	
18	0	0	0	0	0	0	
19	0	0	0	0	1.99	0	8
20	0	0	0	0	0	0	

2. From the data matrix (Messolongi_diatom_density), click **Pre-treatment** > **Transform(overall)...** and choose Square-root.

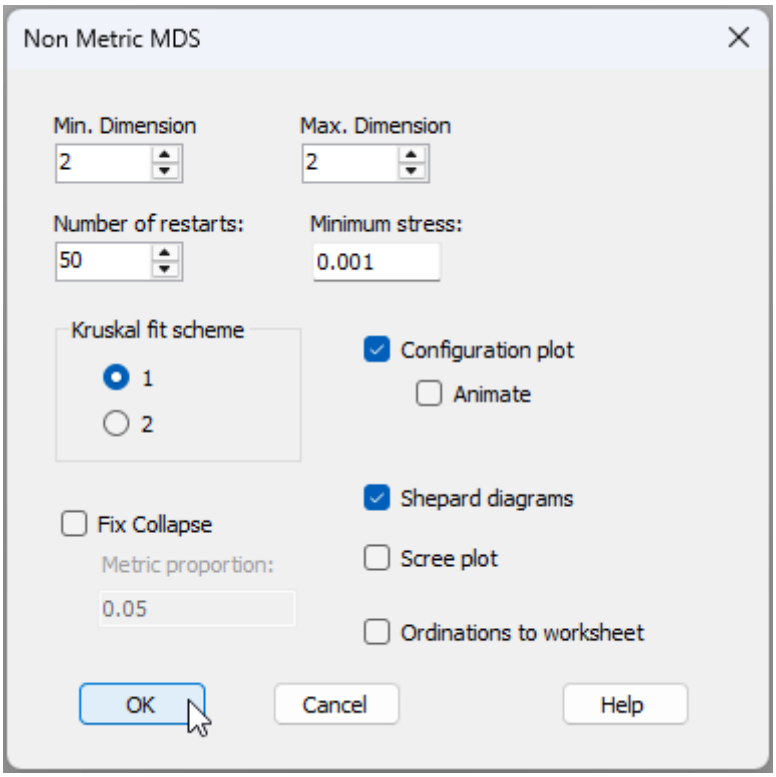


3. The square-root transformed data are held in the sheet named 'Data1'. From this, click **Analyse** > **Resemblance...** and choose to calculate Bray-Curtis resemblances among samples, like so:

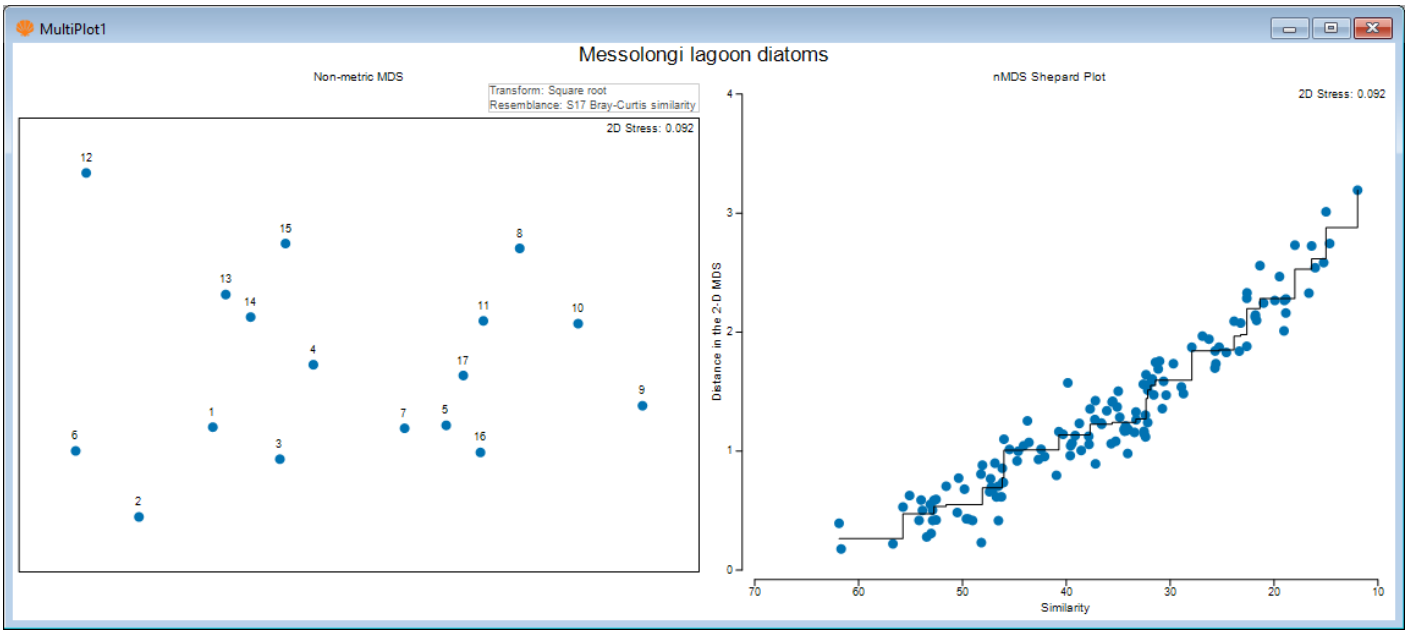


Non-metric MDS (nMDS)

4. The Bray-Curtis similarities are held in the sheet called 'Resem1'. From this, click **Analyse** > **MDS** > **Non-metric MDS (nMDS)...**. Change the 'Max. Dimension' from 3 to 2 (for this exercise, we shall focus only on the 2-d solution), then click 'OK', as shown in the dialog below.

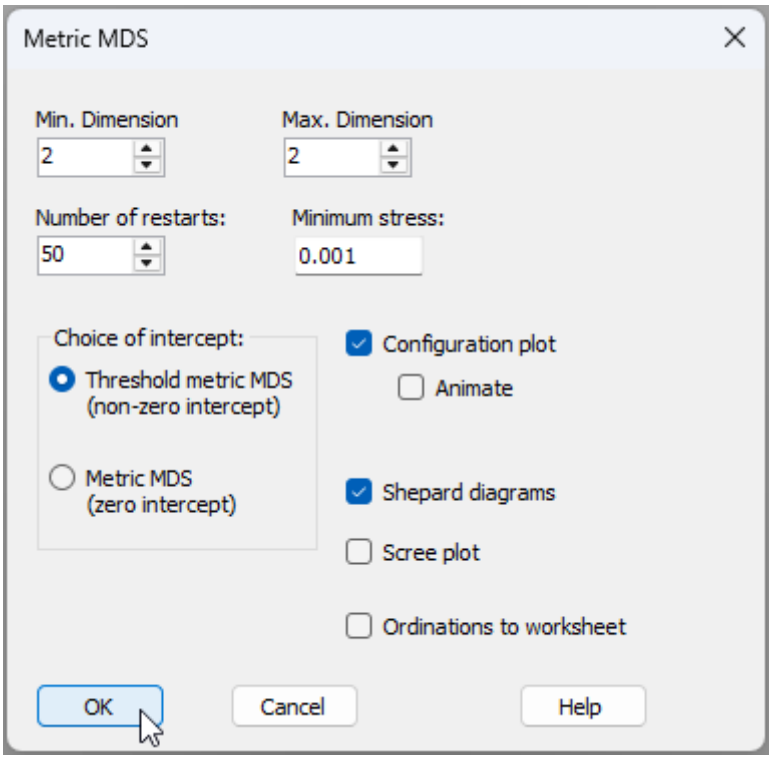


In the results ('MultiPlot1'), we can see the nMDS ordination ('Graph1'), the *monotonic* relationship shown in the Shepard diagram ('Graph2'), and that the value of stress is 0.092.

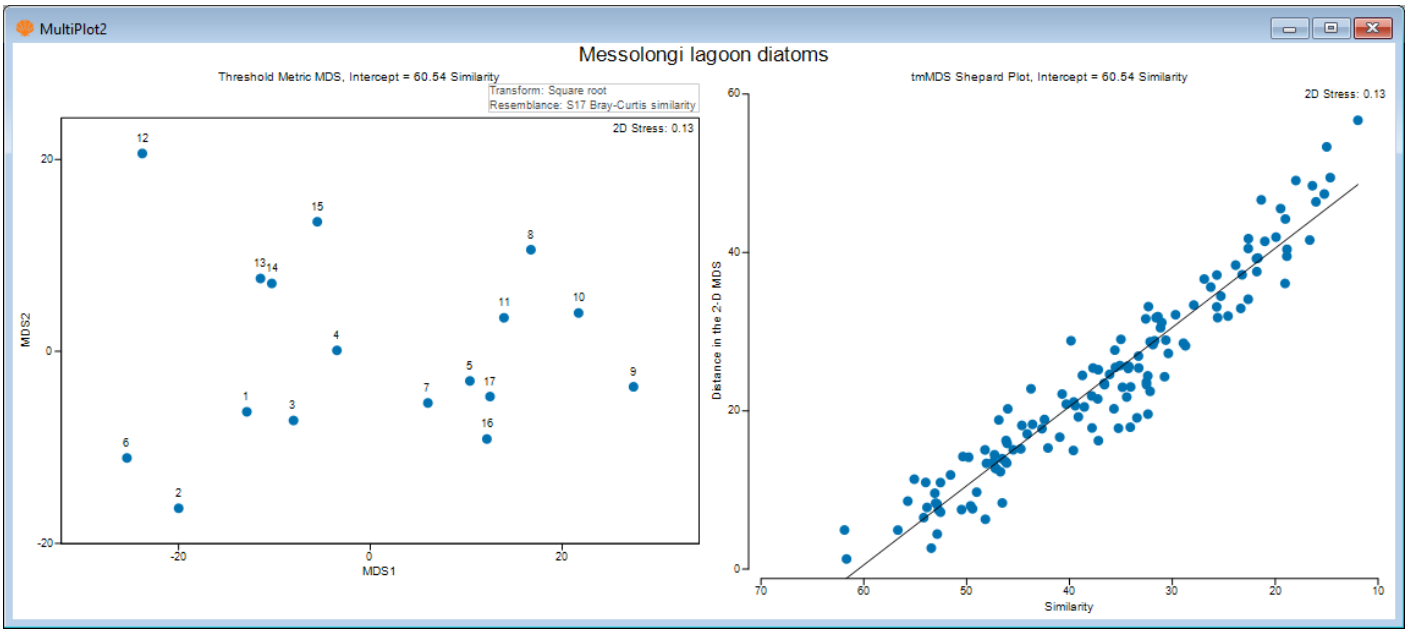


Threshold metric MDS (tmMDS)

5. From 'Resem1', click **Analyse** > **MDS** > **Metric MDS (mMDS / tmMDS)...**, then choose 'Choice of intercept: $\bullet$Threshold metric MDS (non-zero intercept)', change the 'Max. Dimension' from 3 to 2, then click 'OK' (see below).



In the results ('MultiPlot2'), we can see the tmMDS ordination ('Graph3'), the *linear* relationship shown in the Shepard diagram ('Graph4'), with a *non-zero intercept* of 60.54% similarity, and a stress value of 0.130. This is a bit larger than the stress obtained for the nMDS (0.092), but is still sufficiently low (< 0.20) to permit useful interpretations of the patterns of inter-point relationships seen in the diagram.



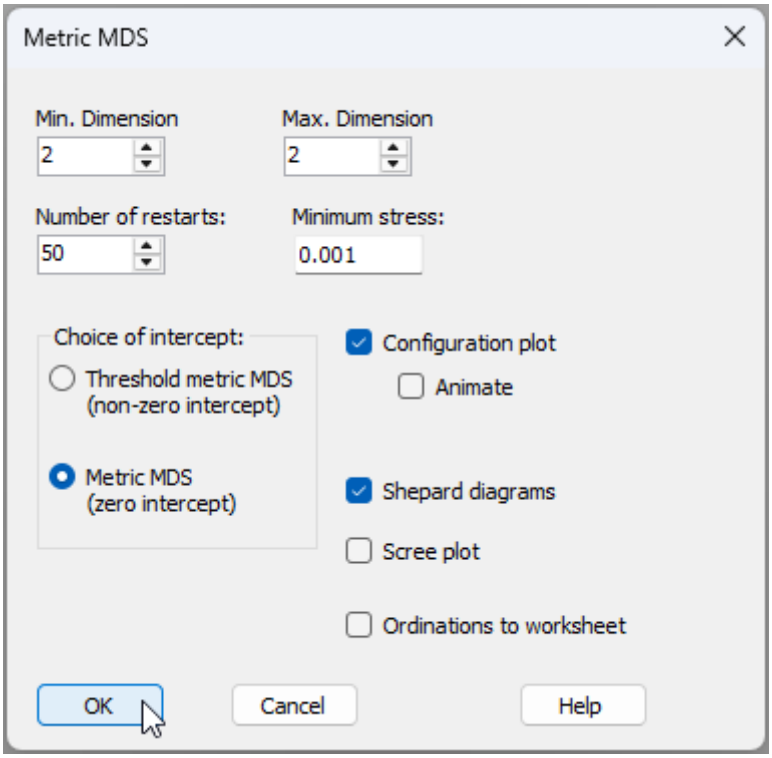
Note that the threshold (intercept) value of 60.54% similarity indicates that two points that are 'on top of one another' in the tmMDS configuration are to be interpreted as being *ca*. 39.46% dissimilar, and not 0% dissimilar.

Also, unlike the nMDS plot (which has no labels on its axes and from which only rank-order relationships among the points can be interpreted), the *axes are labeled* on the tmMDS plot, with values given in the units of the original (Bray-Curtis) resemblance measure. Thus, two points that

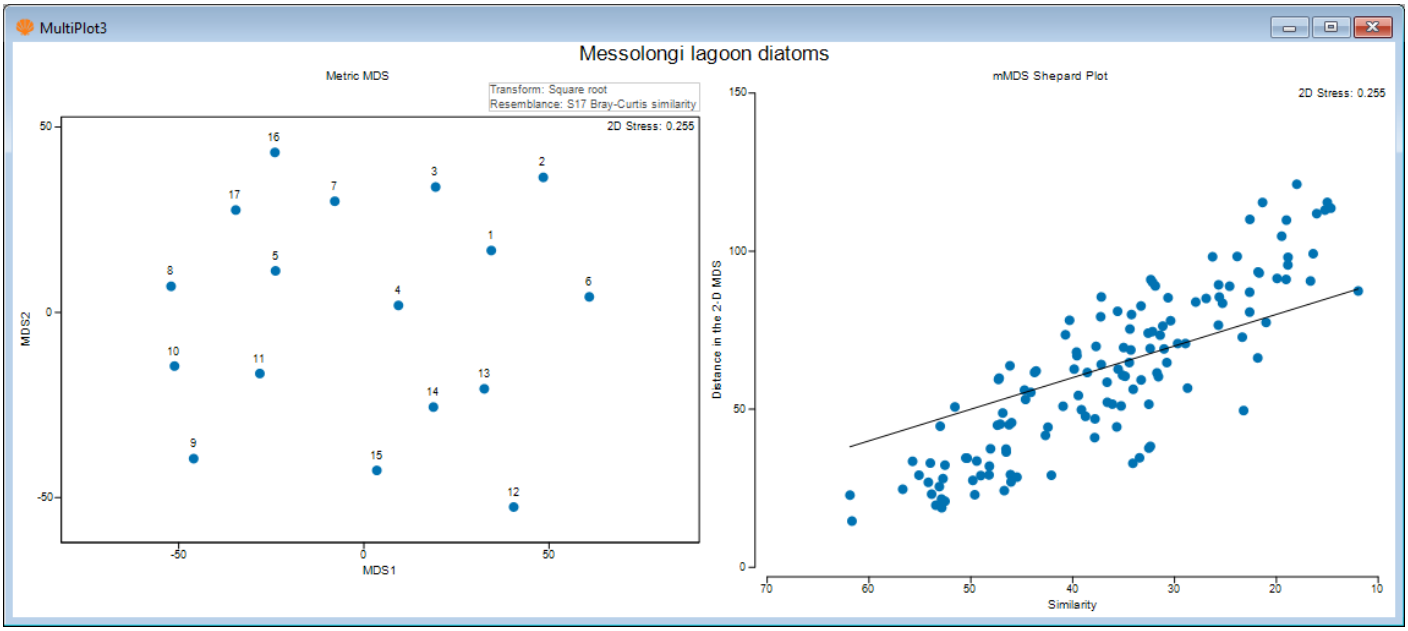
are (say) 20 units apart from one another on this plot can be interpreted as being approximately (20% + 39.46%) = 49.46% dissimilar. Thus, although the tmMDS comes at the price of higher stress compared to the nMDS, it also 'gives something back' in the form of added interpretability (i.e., with labels on the axes permitting the estimation of actual inter-point dissimilarities). So, clearly, tmMDS can be a good option, provided the stress still remains low enough to yield a plot that is adequate for interpretation.

Metric MDS (mMDS)

6. From 'Resem1', click **Analyse** > **MDS** > **Metric MDS (mMDS / tmMDS)...**, then choose 'Choice of intercept: $\bullet$Metric MDS (zero intercept)', change the 'Max. Dimension' from 3 to 2, then click 'OK', viz:

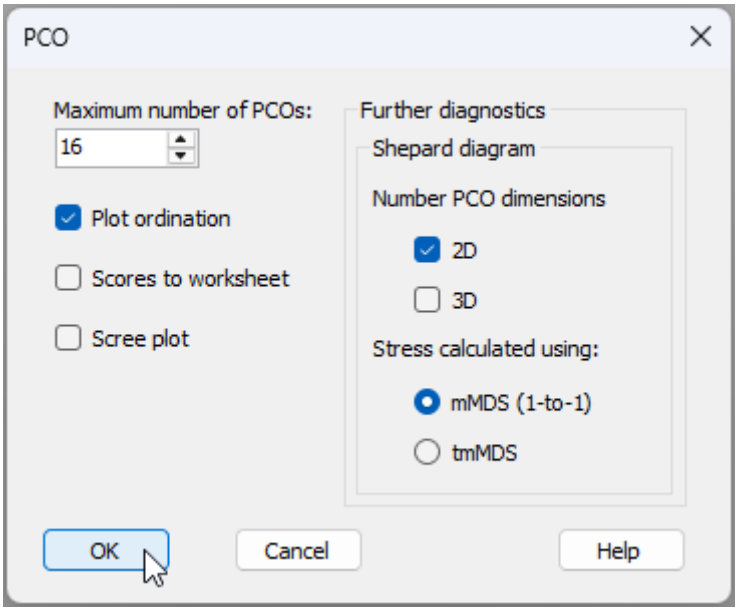


In the results ('MultiPlot3'), we can see the mMDS ordination ('Graph5'), the *linear* relationship with a *zero intercept* in the Shepard diagram ('Graph6'), and a stress value of 0.252. This is much larger than that of either the nMDS (0.092) or the tmMDS (0.130), and is so high (> 0.25) as to prevent meaningful interpretation of any patterns seen in the resulting diagram.

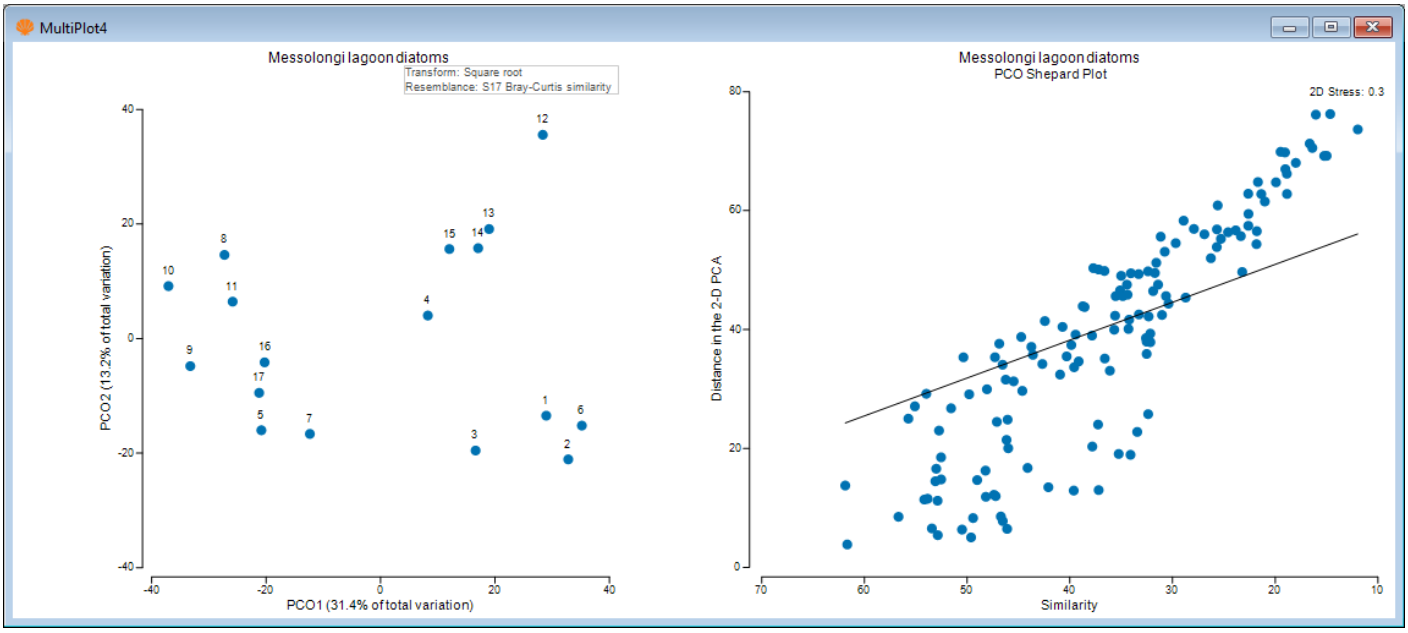


Principal Co-ordinate Analysis (PCO)

7. First, we will run the PCO and get diagnostics for a Shepard diagram and associated stress calculation that uses a zero intercept (as in mMDS). From 'Resem1', click **PERMANOVA+ > PCO**, then choose 'Shepard diagram > Stress calculated using: $\bullet$mMDS (1-to-1)', then click 'OK' (see the dialog below).

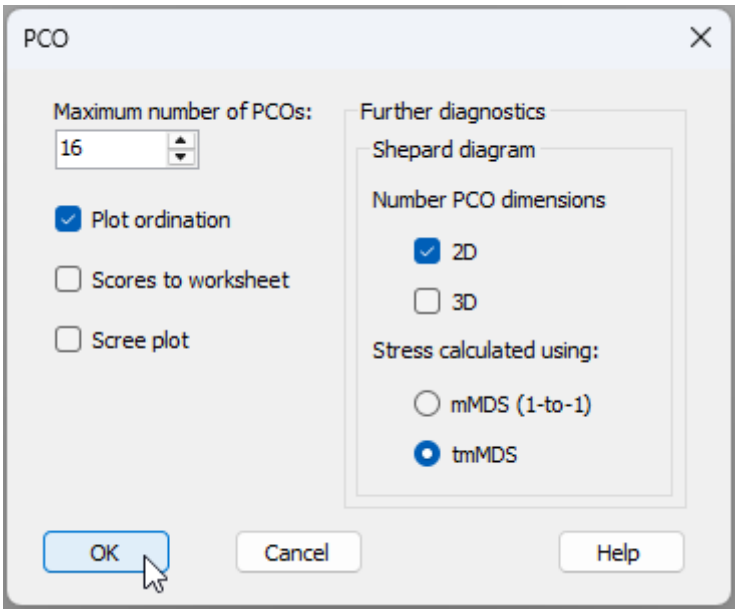


The resulting PCO plot ('Graph7') and associated diagnostic Shepard diagram ('Graph8') are shown in 'MultiPlot4' in the Explorer tree.

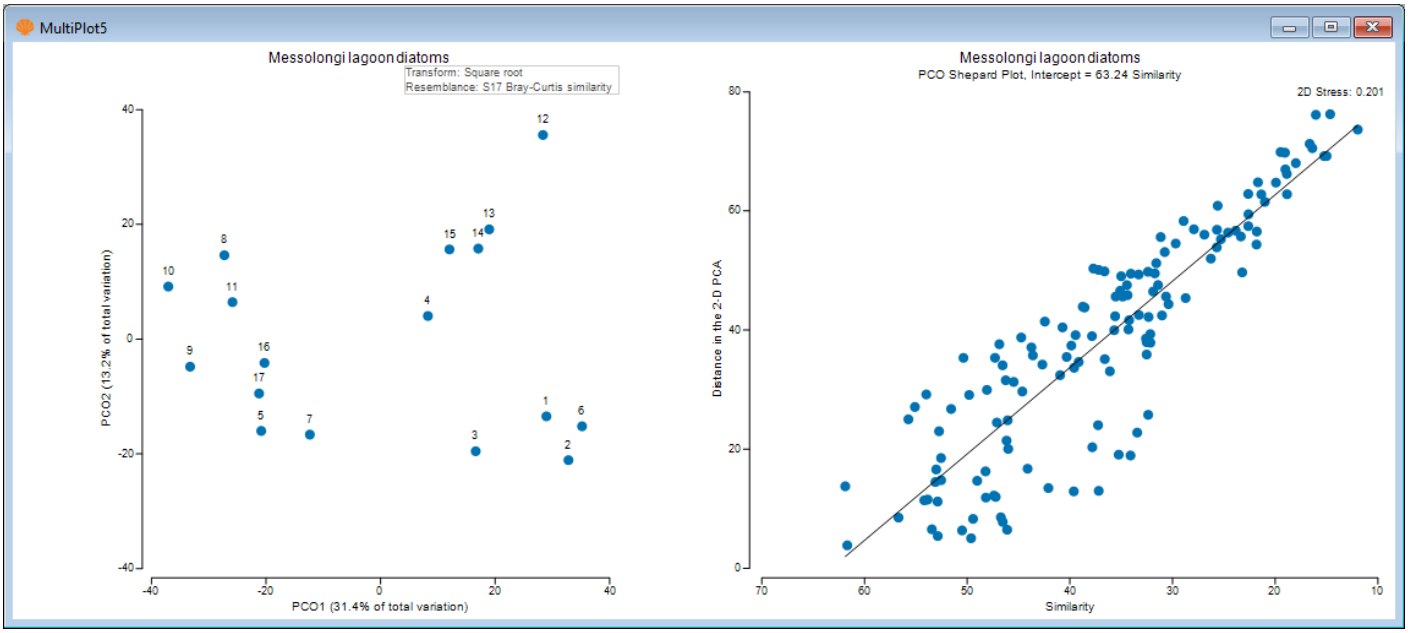


The first two principal axes explain 44.56% of the total variation (less than half, as seen in the output file called 'PCO1' in the Explorer tree). In addition, the stress is extremely high, at 0.30, indicating that many of the inter-point distances seen in the PCO plot do not reflect their true dissimilarity well at all. This is especially true for the small to middle-sized dissimilarities that range from about 40% to 70% (see the large scatter of points in 'Graph8' for values having ~ 60% to 30% similarity along the x-axis).

8. Next, we will run the PCO again, but this time choose 'Shepard diagram > Stress calculated using: $\bullet$tmMDS' (see the dialog below).



The results are given in 'MultiPlot5' in the Explorer tree.



It is clear that the PCO plot itself remains exactly the same (compare 'Graph9' with 'Graph7'). However, we are now using a different yardstick to measure the stress associated with this plot. Specifically, we are permitting the intercept to float away from the origin, and asserting (as we did with the tmMDS) that there is a *threshold dissimilarity* that a pair of samples must exceed before they will occupy two different positions on the configuration plot. Here, the intercept is a similarity of 63.24%, corresponding to a dissimilarity threshold of 36.76%. In other words, any two samples that appear on top of each other on the plot (at a distance of zero) can be estimated to be 63.24% similar, not 100% similar.

By adding this flexibility, we can see that the stress improves dramatically, dropping to 0.201, which borders on becoming acceptably interpretable, at least for the broader structures. Once again, it is the larger dissimilarities that are captured better by the PCO (see the tighter clustering of points around the line-of-best-fit in the upper right area of the Shepard diagram), compared to the smaller ones.

Observations & recommendations

This example serves to demonstrate a more general point. Namely, for a given dataset, the ordering of methods with respect to the value of stress associated with the final configuration will almost always be: nMDS < tmMDS < mMDS < PCO. This is so because:

- all MDS solutions will, by their very design, achieve lower stress than a projection-type ordination, such as PCO, as they are explicitly geared to minimise stress; and
- within the MDS family of methods (nMDS, tmMDS and mMDS), nMDS has the fewest constraints, requiring only a monotonic relationship, while mMDS has the greatest constraints, requiring not just a linear relationship but also a zero-intercept in the Shepard diagram. This is more difficult to achieve, resulting in a higher stress.

Our general recommendations for ordination are:

- Use nMDS, as starting point, to achieve the lowest-stress solution.

- If the Shepard diagram for the nMDS shows that the relationship between configuration distances and original distances is approximately linear, then tmMDS can be used to achieve an ordination with the added benefit of having axes which (along with the added threshold dissimilarity) can be used to estimate actual inter-point dissimilarities directly from the plot. Only retain the tmMDS, however, if the stress is sufficiently low to permit interpretability (< 0.20 as a rule-of-thumb).
- Only go one step further to consider using mMDS if the intercept is sufficiently close to zero to warrant this. (This will be evident in the Shepard diagram).
- PCO will typically *not* be the best tool to use, in general, to visualise inter-point relationships in an ordination, but *will* extract the axis (or axes) of greatest total variation through the data cloud as a whole. This means that large dissimilarities will typically be represented much better than small dissimilarities in a PCO plot.

Utility of PCO

Although PCO may not be the tool of choice for ordination, it is important to recognise that it is still a very useful tool for other reasons. Specifically, a variety of PERMANOVA+ routines use PCO ('under the hood') in order to calculate correctly (using the *full* set of PCO axes, and carefully accounting for negative eigenvalues):

- Distances among centroids (in the space of a chosen resemblance measure);
- Distances from individual points to their own group centroid (in PERMDISP); and
- Monte Carlo p-values (in PERMANOVA).

PCO axes are also used (obviously!) to perform canonical analysis of principal co-ordinates (CAP).

¶*It is sometimes stated that PCO and metric MDS are the same thing. This is simply incorrect and the example here clearly demonstrates this. PCO is a projection onto principal axes, whereas metric MDS is distance-preserving and minimises stress for a configuration drawn in a chosen number of dimensions. PCO may sometimes be referred to as 'classical scaling', but it should not be confused with multi-dimensional scaling (MDS), metric or otherwise.*