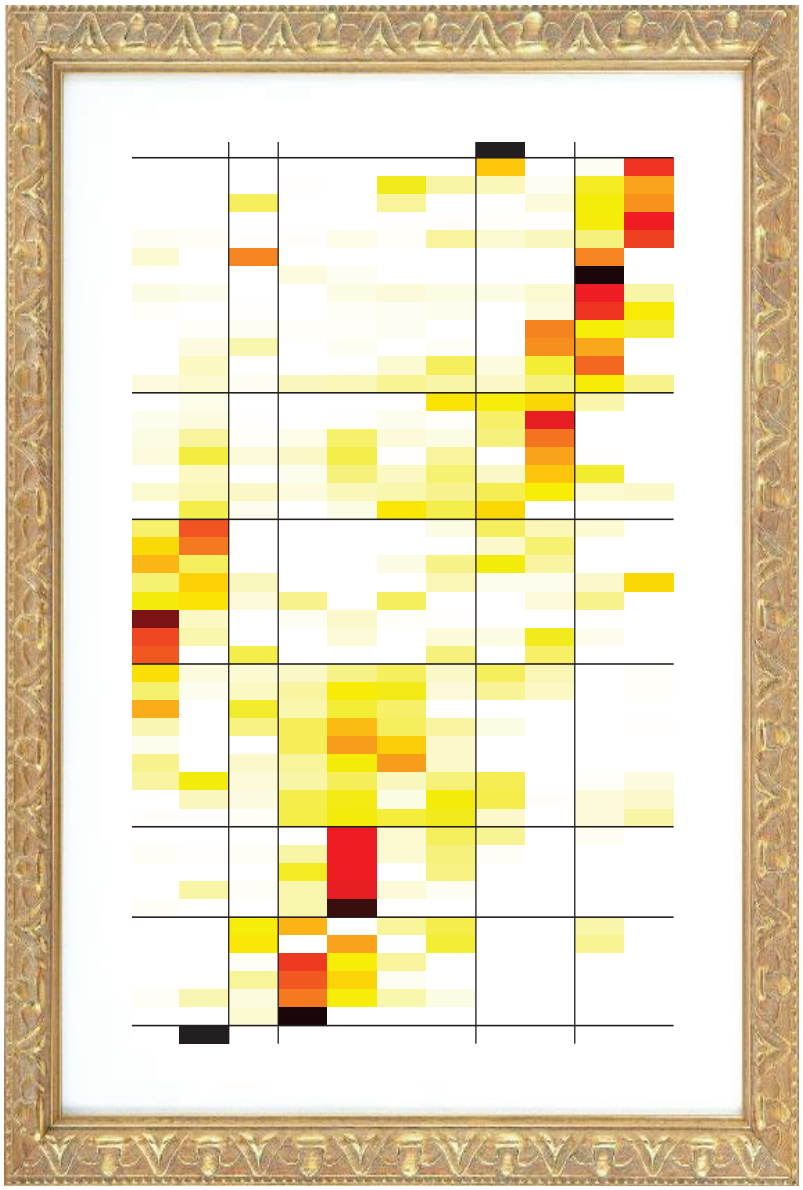
PRIMER v7:

User Manual / Tutorial



K R Clarke & R N Gorley

PRIMER-e
empowering research

***P**lymouth*
***R**outines*
***I**n*
***M**ultivariate*
***E**cological*
***R**esearch*

PRIMER v7:
User Manual/Tutorial

K R Clarke & R N Gorley

Published 2015
by
PRIMER-E Ltd
Business Address: 3 Meadow View, Lutton, Ivybridge, Devon PL21 9RH, United Kingdom

Reproduced from 15 August 2016
by
PRIMER-e (Quest Research Limited)
Business Address: ecentre, Gate 5 Oaklands Rd, Massey University Albany Campus, Auckland, New Zealand

First edition 2015

Clarke, K.R., Gorley, R.N. 2015
PRIMER v7: User Manual/Tutorial
PRIMER-E: Plymouth

PRIMER v7: User Manual/Tutorial

Contents

OVERVIEW

MANUAL/TUTORIAL

OVERVIEW

A. Contact details and installation of the PRIMER v7 software

Getting in touch with us

For any up-to-date news about PRIMER, including details of upcoming PRIMER workshops, see our web site at: *http://www.primer-e.com*

Please report any bugs or technical problems to: *tech@primer-e.com*

For licensing and other general enquiries, contact the PRIMER-e office at: *primer@primer-e.com*

Our business postal address is:
PRIMER-e (*Quest Research Limited*)
ecentre
Gate 5, Oaklands Road
Massey University Albany Campus
Auckland 0632
New Zealand
Tel: +64 (0)9 869 2230

(or you may, if you prefer, use the registered address of the company: PRIMER-e (*Quest Research Limited*), 67 Mahoenui Valley Road, RD3 Albany, Auckland 0793, New Zealand)

System requirements

PC with Intel compatible processor.
Any modern Windows operating system (XP or later).
Sufficient RAM memory to run the operating system satisfactorily. More memory will be required for very large data files.
A PDF reader program such as Adobe Acrobat Reader for reading the manuals.
Optionally Excel 2000 or later installed on the PC if you want to read and save Excel data files.

Installing PRIMER v7

You need to be logged on as an administrator. You can keep version 6 or earlier versions installed. Download the latest setup file from our web site. Install by double clicking on the file. (If installing the trial version off-line then you must have a reasonably modern version of the Microsoft .Net framework installed). PRIMER v7 will run in trial mode for 30+ days, but eventually needs to be activated (authenticated) whilst on-line, using a licence key purchased from PRIMER-e.

Information on analyses

Detailed information and examples of virtually all the analyses offered by PRIMER v7 are found in the accompanying Methods manual: Clarke KR, Gorley RN, Somerfield PJ, Warwick RM 2014, *Change in Marine Communities, 3rd edn*, PRIMER-E Ltd, Plymouth (henceforth referred to by its initials, CiMC). However, in order not to break the flow of the coherent strategy outlined there, in a few cases it was preferable to give more detailed formulae and descriptions here (e.g. of all the resemblance measures available, and the necessary adjustments to each in the case of missing data etc). In addition, the software has a Help system, but the availability of comprehensive Methods and User/Tutorial manuals in searchable pdf format, downloadable from the Help menu, has allowed the Help system itself to be relatively succinct.

PERMANOVA + add-on

In examples in this manual you may see the main menu item PERMANOVA+. This is an add-on product to PRIMER, operating with PRIMER 7 exactly as it did with PRIMER 6. Its operation is covered by another (combined) Methods/User manual: Anderson MJ, Gorley RN, Clarke KR 2008, *PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods*, PRIMER-E, Plymouth.

If you are familiar with PRIMER 6, it is suggested you read Section C, on the many enhancements in PRIMER 7, and then try out the software on some of the Examples that come with the software, available from the PRIMER Help menu (under Get Examples V7). New features in version 7 are indicated in the manual by a vertical red dotted line in the margin.

New users should start with Sections B, D and E, and then work selectively through the detailed material of the manual. This is written in the style of a continuous Tutorial but functions also as detailed Reference material. Analyses of specific data examples can be tracked via the 'Index to data sets' at the end of the manual.

B. Introduction to the methods of PRIMER

Application areas

PRIMER 7 (Plymouth Routines In Multivariate Ecological Research) consists primarily of a wide range of univariate, graphical and multivariate routines for analysing arrays of species-by-samples data from community ecology. Data are typically of abundance, biomass, % area (or line) cover, presence/absence etc, and arise in biological monitoring of environmental impact and more fundamental studies, e.g. of dietary composition. Also catered for are matrices of physical values and chemical concentrations, which are analysed in their own right or in parallel with biological assemblage data, 'explaining' community structure by physico-chemical conditions. The methods of this package make few, if any, assumptions about the form of the data (non-metric ordination and permutation tests are fundamental to the approach) and concentrate on approaches that are straightforward to explain. This robustness makes them widely applicable, leading to greater confidence in interpretation, and the transparency possibly explains why they have been adopted worldwide, particularly in marine science but also in terrestrial and freshwater ecology, forestry, soil science etc. The statistical methods underlying the software are explained in non-mathematical terms in the accompanying methods manual (Change in Marine Communities, 3rd edition, 2014), which also shows outcomes from many literature studies, e.g. of environmental effects of oil spills, drilling mud disposal and sewage pollution on soft-sediment benthic assemblages, disturbance or climatic effects on coral reef composition or fish communities, more fundamental biodiversity and community ecology patterns, mesocosm studies with multi-species outcomes etc. Many of the data sets used in the methods manual (abbreviated to CiMC), and all of those used in this User Manual/ Tutorial are available with the installation so that the user can replicate the analyses.

Though the analysis requirements for biological assemblage data are a principal focus, the package is equally applicable (and increasingly being applied) to other data structures which are either multivariate or can be treated as such. These include: multiple biomarkers in ecotoxicology, and their relation to water or tissue concentrations of chemical contaminants; composition of substrate in geology or materials science; morphometric measurements in taxonomy; genetic studies and especially microbial analyses of large numbers of OTUs; signals at multiple wavelengths in remote sensing; even environmental economics, state variables in complex mathematical box models, acute medicine, epidemiology, etc. Univariate measurements which can sometimes be treated more effectively in a multivariate way include particle size analysis for water or sediment samples and size frequency distributions of organisms in cohort studies (the multivariate variables are the discrete particle or organism size classes). Sets of growth curves for individual organisms tracked through time (repeated measures, thus correlated) can also be handled. The unifying feature is that all data sets are reduced to an appropriate triangular matrix representing the resemblance of every pair of samples, in terms of their assemblages, suites of biomarkers, particle size distributions, shape of growth curves, etc. Clustering and ordination techniques are then able to display the relationships among the samples, and permutation tests impose a necessary hypothesis testing structure. To demonstrate the range of application areas, references to many thousand publications citing PRIMER software can be downloaded from the PRIMER-e website (www.primer-e.com).

Basic routines

The routines of the package cover: data pre-treatment (transforms, dispersion- and other variable-weighting, assessed by Shade plots); about 50 resemblance measures (now allowing missing data); hierarchical clustering of samples (or species) via standard agglomerative and novel divisive and 'flat' techniques; ordination by non-metric (*n*MDS) and metric multidimensional scaling (*m*MDS, *tm*MDS), and principal components (PCA), to summarise patterns in biotic and abiotic samples; permutation-based hypothesis testing (ANOSIM, also in novel ordered form), testing *a priori* group structures of multivariate samples, from different times/locations/treatments etc; a strong emphasis now on species patterns (novel Coherence curves and Shade plots); linking of multivariate biotic patterns to suites of environmental data or other biotic arrays (BEST, LINKTREE); comparative (Mantel-type) tests of similarities to model structures (RELATE, including novel 2-way forms); second-stage analyses (2STAGE) for 'repeated measures' and comparison of analysis choices (of taxonomic level, transform, resemblance coefficient); suites of diversity indices, dominance plots, SAD curves, species accumulation estimators, taxonomic aggregation etc, and tests for biodiversity indices based on taxonomic distinctness of species (TAXDTEST); novel region estimates for mean communities from multivariate bootstrapping; and a wide range of other data manipulation and graph types (bar, line, mean, box, scatter, surface, histogram and shade plots), new to PRIMER 7.

C. Changes from PRIMER 6 to PRIMER 7

Wizards and major new analysis options

Dialog boxes generally are more 'wizard'-based, unifying and simplifying parameter input and, importantly, there is a new **Wizards** menu (Section 10) with three items, as follows:

1) Basic multivariate analysis wizard which gathers together the main steps of a core PRIMER analysis in a single dialog box. It is aimed at new users who may initially have trouble formulating a reasonable choice of a sequence of routines. It covers pre-treatment options such as standardising and transforming samples, calculating resemblances, running group-average clustering, non-metric MDS, SIMPER, and either ANOSIM or (Type 1) SIMPROF for testing *a priori* or *a posteriori* groups, and offers robust defaults, tailored to data type (biotic or environmental) and availability of a factor with repeat levels. Importantly, the analysis sequence is fully laid out in the Explorer tree, allowing the new user to reconstruct the menu selections needed for this basic analysis.

2) Matrix display wizard which runs the new and comprehensive **Shade Plot** routine (Sections 4 & 10) which is an 'image' of the data matrix, with species abundances – or other quantity values – shown by depth/colour of shading. At a simple level, shade plots aid choice of transformation or other data pre-treatment but, when the species and/or sample axes are suitably grouped, clustered, seriated, non-linearly linked or subject to combinations of these, they become powerful tools for interpreting sample patterns (established by testing and seen in ordinations) in terms of individual species driving those patterns. The Matrix display wizard sets up the (rather involved) sequence of sample and species resemblance calculations, and clustering and seriation steps, to give a robust, initial shade plot, which the user can then refine by further ordering or constraining of axes.

3) Coherence plots wizard which uses a novel (Type 3) SIMPROF testing series on standardised or normalised biotic or environmental variables, to define e.g. groups of Coherent species having statistically indistinguishable patterns of response over the samples, within sets, and statistically different responses among sets. The wizard then displays the sets using the new **Line Plot** routine.

4) The **CLUSTER** menu now includes a new linkage option under its agglomerative hierarchical menu, in addition to UPGMA, single and complete linkage, namely flexible beta – a standard WPGMA extension – and a cophenetic distance matrix can be output to allow computation of cophenetic correlation (Section 6). More novel, however, is the introduction of the following new divisive and 'flat' clustering methods, designed for the non-parametric PRIMER framework:

5) UNCTREE, a binary divisive algorithm which is an unconstrained form of the v6 **LINKTREE** routine, e.g. successively dividing groups so as to maximise the ANOSIM R statistic between the two groups so formed. SIMPROF (Type 1 or 3) tests give a stopping rule for the binary divisions.

6) kRCLUSTER, a flat-form (non-hierarchical) method, based on the idea of *k*-means clustering in which the full set of samples are divided into a pre-specified number (*k*) of groups, minimising the within-group sums of squares (≡ within-group squared Euclidean distances). Generalising this to PRIMER's context, kRCLUSTER seeks to find a division into *k* groups which maximises the non-parametric ANOSIM R statistic, and is therefore definable for any resemblance measure, not just Euclidean distance. The routine is also able to choose *k* by computing *k*-R clusters for successively larger *k* until none of the groups is statistically heterogeneous, as seen by (Type 1 or 3) SIMPROF.

7) The **MDS** menu for non-metric MDS (***n*MDS**) now allows calculation of ordination axes in any number of dimensions (2 or more), not just 2 and 3, and any combination of the higher-d axes can be plotted in a 2- or 3-d plot (as for PCA in v6); a *Scree plot* shows the declining MDS stress for increasing dimensionality. Another important addition to *n*MDS is the ability to Fix collapse of the non-metric MDS plot when a sample (or samples) are sufficiently distant from remaining samples to cause an indeterminacy in the rank order information. This novel procedure is implemented by using both of the ideas in the following new additions to the MDS routine (Section 8):

8) Metric MDS (***m*MDS**) seeks to preserve the actual dissimilarities in the resemblance matrix as distances in the low-d ordination, rather than preserving only their rank order (as in nMDS). This can be a very successful alternative to *n*MDS when there are very few points (perhaps 4 or 5) and the rank orders do not carry enough information; this can happen easily for plots of group means. (Note, this is not PCO, as implemented in PERMANOVA+, which projects points into low-d space, whereas *m*MDS places points in that space, minimising stress in the linear Shepard plot).

9) Threshold metric MDS (*tm*MDS) which, instead of fitting the Shepard plot by a straight line through the origin (*m*MDS), fits a straight line with an intercept. It borrows from *n*MDS the ability

to truncate small dissimilarities to effectively zero distances in the ordination (reflecting the fact that replicates from exactly the same condition are never 0% dissimilar because of sampling error) and from *m*MDS the preservation of linear distance additions with linear dissimilarity changes, where conditions differ. It can be a useful compromise for low species turnover among samples.

10) Combined MDS minimises an equal mixture of stress functions from two *n*MDS ordinations, which has potential application to produce a consensus view of among-sample relationships for two sets of variables which cannot be merged into a single matrix – perhaps needing different resemblance measures (e.g. biotic and abiotic; motile organism counts and colonial species areas). It is combined stress from *n*MDS with a small *m*MDS component that 'fixes' an *n*MDS collapse.

11) The **Bootstrap Averages** routine is another significant innovation in v7, providing a region estimate for each group mean in a 2-d (or 3-d) ordination plot from samples with an *a priori* one-way group structure (or a 2- or higher-way crossed group design flattened to simple groups with replicates). This bootstraps the samples in an m-dimensional *m*MDS space for which m is large enough for the among-sample distances to closely match the original dissimilarities (as judged by Pearson matrix correlation > 0.99 say) but small enough to avoid the unrepresentativeness of bootstrap samples in very high dimensions. The averages of repeated bootstrap samples for each group are ordinated into 2- or 3-d to form a region estimate for mean communities, which (in 2-d) is smoothed and marginally bias-corrected (but not formally a confidence region), Section 17.

12) The **ANOSIM** routine has been greatly expanded, firstly to include a concept of ordered group structure for one (or more) of the factors input to the ANOSIM test. This tests the null hypothesis of no group differences against a directed alternative in which the groups are in specified sequence (e.g. years under a time trend, spatial gradients, increasing impact conditions). It then permits a more powerful test, based on generalising the ANOSIM R statistic to an ordered R^O (the slope of a regression of dissimilarity ranks against the ranks of a model 'seriation with replication' matrix) – a test which can also be run in the absence of replication, Section 9.

13) A second major extension of **ANOSIM** is to designs with three factors (A,B,C) in all feasible crossed and nested combinations – fully crossed A×B×C, fully nested C(B(A)), nested in crossed C(B×A), and crossed with nested B×C(A). All cases allow any factor to be ordered or not, and non-replicated models either exploit ordering or extensions of the previous approach to 2-way crossed designs without replication (e.g. inferring a B effect from commonality of B level patterns across the levels of A, as measured by a non-parametric matrix correlation ρ), Section 9.

14) The **RELATE** routine is extended (Section 14) to include a 2-way RELATE test, operating in a similar way to 2-way crossed ANOSIM. That is, the matrix-correlation matching statistic (ρ, now including Pearson as well as Spearman and Kendall) of the resemblances to any model (or biotic) matrix is calculated within the strata of a second (group) factor, and averaged. The permutations for the test are now similarly constrained within those strata, so that any effects of a second factor (e.g. site differences) are removed from the test of the first (e.g. an annual trend or seasonal cyclicity).

15) The **BEST** routine is similarly extended to a 2-way BEST procedure and 2-way BEST test, by choosing explanatory (e.g. environmental) variables which 'best explain' the multivariate pattern in the response (e.g. community) variables, simultaneously within strata of a secondary ('nuisance') factor, Section 13. E.g. this can remove location differences in base communities among oil-fields, when fitting contaminant variables to community patterns, simultaneously around several oil-fields.

16) The **SIMPROF** test (now Type 1 SIMPROF) for multivariate structure in (subsets) of samples, mainly used with agglomerative (CLUSTER) and constrained divisive clustering (LINKTREE), has been generalised to all four combinations of sample or variable resemblances, with permutation over samples or variables. E.g. Type 2 SIMPROF is a test for any association amongst species, and Type 3 is used with a cluster analysis (as with Type 1) to test for heterogeneity in associations over a subset of species, which would allow further sub-division of that subset (Section 10) – this is the core component of the Coherence plots wizard. Use of SIMPROF in CLUSTER and LINKTREE therefore now extends to Type 3 (on variables), and both Type 1 and 3 SIMPROF are also options within the new UNCTREE and kRCLUSTER clustering (Section 6).

17) Another new menu item is **Summary Stats**, a minor feature in analysis terms since it simply computes Min, Max, Average, Sum, SD, Variance, Range and number of Non-zero entries for every variable or for every sample, but it has widespread utility in preparing for other analyses, spotting outliers (run Max in both directions), identifying low occurrence species etc (Section 3).

18) Less noticeable, but with far-reaching ramifications, is that **Resemblance** calculations will all now operate in the presence of Missing! data entries, Section 5. Pairwise elimination of samples (or variables), with one or both entries missing, is undertaken separately for each calculated pair, and each coefficient corrected, where necessary, for their particular bias (arising from unequal numbers of terms in summations). Estimation of missing data by the EM algorithm – the previous **Missing** routine – is still available, and preferable where the rather strict model conditions for its use are met, but the simple bias corrections allow reasonable analyses in other cases with unavoidable and commonly occurring missing data (e.g. samples as questionnaire returns, variables are questions).

Additions to configuration (ordination) plots

PRIMER 7 has added several significant display features (Section 8) to its ordination plots, *n*MDS, *m*MDS and PCA (they also operate with PCO, dbRDA and CAP in the PERMANOVA+ add-on):

19) Bubble plot for a single variable – with values superimposed as circles of differing sizes on the sample points – can now have bubbles of different colours, dependent on the level of a group factor (colours are user-controlled through the usual symbol plotting mechanism).

20) A bubble colour saturation option makes labels plotted on bubbles more visible on a lighter background, and opacity control can make bubbles transparent, making hidden bubbles visible.

21) The key defining bubble sizes is now under user control, allowing specification of the number of bubble sizes drawn in the key, and the data values (actual or as a percentage of their range).

22) Bubble plots for one variable can be drawn with a single user-supplied image, e.g. of a relevant organism (as .jpg, .png etc), displayed at different (rectangular) sizes in place of a (circular) bubble.

23) Bubble plots are now possible in 3-d configurations, utilising a '3-d effect', giving a reasonable facsimile of a 3-d bubble. The 3-d effect can also be selected for 2-d plots, where it can be quite effective in making superimposed (preferably single-character) labels on each point stand out.

24) A new Segmented bubble plot construct is obtained by specifying more than one variable (*k*, say) to display – ideally on 2-d plots though it will operate in 3-d. Circles are divided into *k* equal sectors of different colours, and the sectors plotted at different sizes according to the data values for that point and variable. The colours, and variable order round the circle, are under user control.

25) The existing Spin option to rotate a 3-d configuration can now be captured as an animation file (in .mp4 or animated .gif format), along with any manual interventions which change the angle of view etc. The sampling rate (frames per second) and image size are under user control. This should allow 3-d ordination plot rotations to be embedded in, for example, a Powerpoint presentation or as supplementary material for an on-line publication.

26) Ordination plots in which the points form a natural series (in time or space) can be displayed in animated form, in 2-d or 3-d, with points and/or the joining trajectory (or trajectories) fading in and out in this natural order. Sequence animations can again be captured in an animation file; the speed of traverse through the series is under dynamic user control, and there is initial selection of decay speed, for fade-out of displayed components. It can be used, for example, in tracking natural or impact-induced temporal change (and perhaps recovery) in a longish time series, especially where the community is similar at different times and a static MDS plot of the whole series is cluttered.

27) A third animation option, which can again be captured in .mp4 or .gif format, is the evolution of the MDS iterative process. This is designed as a teaching tool, to see in action how an MDS configuration can sometimes get trapped in a local minimum, and thus the necessity for restarting the iteration from many different initial random configurations of the points.

28) There can now be Split trajectories joining points on an ordination, e.g. multiple time series trajectories drawn for a series of sites in different line types and colours (the latter determined by the symbol colour for the relevant points). Two factors are specified – a numeric factor defining the order in which points are joined and a categorical factor whose levels determine the separate trajectories. [In the authors' experience, this could be one of the most used of the new features!]

29) MDS diagnostics are enhanced by provision of the Minimum Spanning Tree. This is computed for the samples under study as the set of connections of samples to each other, on a single (though branched) route, such that the sum of all the connecting dissimilarities is minimised. It is drawn as a branched trajectory on the 2-d or 3-d MDS plot, so that if it visually departs from an MST that would clearly reduce the total connected length on the low-d plot, this is evidence of stress.

30) An alternative diagnostic is to Join pairs of points with similarity greater than (or dissimilarity less than) some supplied threshold value – in practice a series of threshold values, sequentially –

and look for conflicts in the low-d representation given by the ordination (e.g. points close together but not joined, compared with points further apart but joined, i.e. with lower dissimilarity).

31) An Align option now rotates and reflects (and possibly shrinks/stretches, preserving the aspect ratio) the active configuration to best match another supplied ordination (Procrustes analysis). This could be done manually, and thus less precisely, in v6 but the main advantage here is in simplicity and speed when comparing several ordinations under different transforms, taxonomic levels etc.

32) 2-d ordination plots can be visually merged with a full hierarchical cluster analysis using a Raised dendrogram plot, in which the cluster dendrogram is displayed in the third dimension and the whole structure able to be (manually) rotated, as usual.

Other new plots & plot features

There are many new plot types in PRIMER 7, greatly expanding the ability to view data structures (**Plot** menu, Section 7, but also introduced throughout this manual, wherever they find application).

33) A new *Multiplot* automatically generates groups of related plots together in a single window, to allow broad overview or to use as 'thumbnail' – clicking on individual plots makes them accessible for manipulation. These can all be plots of the same type (eg a sequence of histograms or line plots) or different types (eg MDS ordinations and their Shepard diagrams, in a range of dimensions, and a scree plot). Users can create their own multiplot and fill this with any combination of (single) plots.

34) A major new **Shade Plot** routine images the data matrix, with numerous display combinations for ordering/grouping of both axes. It is the core of the Matrix display wizard – see (2) above – but can be run as a stand-alone routine (e.g. in Section 4 on aiding choice of transformation).

35) Standard **Box Plot**s and **Means Plot**s, for univariate data such as sets of diversity indices, allow respectively the usual non-parametric display of (medians, quartiles, ranges) and normality-based means and 95% confidence intervals for those means, for the supplied group factor of a one-way layout with replication (Section 15). Means plots can use common or separate variance estimates.

36) The histogram plots previously only output as null hypothesis distributions for multivariate tests are now available as a stand-alone **Histogram Plot**, e.g. to use in assessing individual needs for transformation of environmental variables. Histograms for all variables are put into a multiplot.

37) A **Line Plot** displays joined matrix values for a variable (y) along the matrix sample order (x), simultaneously for each variable (with different symbols and joining line colours) in a single plot. Supplying an indicator dividing the variables into groups results in several line plots, held in a multiplot. Such a Line Plot is the main display from the Coherence plots wizard – see (3) above.

38) By default, a **Bar Plot** is stacked, e.g. showing the breakdown of abundances over species for each sample (often for meaned samples then sample-standardised, to give % breakdown), but it can display individual species bars side-by-side in groups for each sample, and has a 3-d form.

39) A **Surface Plot** is relevant only to variables in a meaningful order, e.g. size-classes in particle size distributions or growth curve data, displaying a 3-d surface of (sample, species, data value).

40) Scatter Plot produces a single 2-d (x, y) or 3-d (x, y, z) plot of the sample values from 2 (or 3) specified variables which can be from different worksheets, eg allowing a scatter plot of a diversity index or counts of a single species against an abiotic PC or single variable. Points can be labelled and a group factor can be used to give differing symbols/colours of points for the group levels.

41) An existing plot with a new display structure is the **LINKTREE** tree diagram, now by default more like a CLUSTER dendrogram, which greatly aids flexible identification of samples by labels or symbols – rather than just sample numbers – in the resulting SIMPROF groups, for example. The previous format is retained as an optional 'classic' layout. Also, the y axis can use equi-stepped binary divisions (A% scale) in addition to the previous scale using size of group separation (B%).

42) PRIMER 7 also adds a number of significant additional features to existing plots:

a) The general Graph Options dialog applying to all plots and which controls display of titles, labels, symbols, axis scales, etc now adds a tab for Variable symbols and labelling (e.g. in Shade Plots), and a new Key tab to allow control of key label/title sizes, selective suppression of keys etc.

b) As for the Scatter Plot, a **Draftsman Plot** is now able to utilise a group factor to allow different symbol shapes/colours for the different levels of that factor, across the whole plot.

c) Individual points on a Shepard diagram from an MDS run may now be clicked on, to display the two sample labels this point represents. This can aid identification of outliers – samples which fit poorly into the low-d space. The dimension and stress value is now shown on the Shepard diagram.

d) **Cancel Zoom** is a new Graph menu item (and icon on the Tool Bar) to allow a quick return to the unmagnified plot; similar is a **Reset** option on the Graph menu which (in addition to cancelling a zoom) will, for example, restore an MDS plot to its original orientation before a manual rotation.

e) A Monochrome option on the General tab turns all colour displays to mono, and replaces colour fills with monochrome hatching, patterns for which can be chosen in the Key which defines colour.

f) Key dialogs, of whatever type, can be instantly accessed by clicking on the key in the plot. This extends to some other plot 'hot' areas – clicking on titles or axes brings up the relevant dialog.

General & miscellany

43) PRIMER 7 is now a downloadable product, with the same download serving for trial, licenced and update purposes (with/without the PERMANOVA+ add-on). Full functionality requires the relevant key to be purchased, and needs authentication automatically via an internet connection (of course the software will then operate off-line). There is an off-line authentication process, but this should be avoided if at all possible – it needs manual code exchanges with the PRIMER-e office.

44) Maintenance updates to v7 will be offered automatically on release, the next time PRIMER is opened and there is an internet connection – these are quickly implemented. This facility can be switched off by **Options** (under **Tools)**, and a manual check made for updates via the **Help** menu.

45) PRIMER 7 now has an **Undo** facility on the **Edit** menu so that any changes to data entries can be (multiply) reversed – this does not apply to operations on other menus which create new sheets (e.g. **Pre-treatment**) since these can always be deleted in the Explorer tree and re-run. An **Undo workspace** on the **File** menu reverses deletions, name changes etc on the Explorer tree.

46) At PRIMER-e, we are not keen on upgrades which simply 'shuffle the furniture' but a small number of routines have been moved to different menu positions, for important reasons.

a) **Pre-treatment** becomes a main menu in its own right, since it is an almost inevitable first step. **Transform (individual)** has logically been moved into this menu from the **Tools** menu.

b) The active sheet for **BEST** is now the (usually biotic) resemblance matrix not, as previously, the explanatory (usually abiotic) variables sheet. This is because the resemblance matrix defines the specific 'response' samples to be analysed and the explanatory variables could be from a larger look-up table covering environmental data for a larger region or time. This change also makes the logical link with DISTLM in the PERMANOVA+ software – a more precise analogue of multiple linear regression which, as with all the add-on routines, has the resemblance matrix as active sheet.

c) The same reason of defining samples to be analysed as those in the (biotic) resemblance matrix makes this the active sheet for **LINKTREE**, which has logically been moved under **CLUSTER** – it is a constrained version of the unconstrained **UNCTREE**, which must start from a resemblance.

47) A more flexible variable information worksheet replaces the old aggregation file (which can store trait and score information in addition to taxonomy), e.g. species can now be selected, ordered etc using indicators on variable information. This is significant for taxonomic distinctness work, as is the alternative entry of species similarity matrices to **TAXDTEST** and **DIVERSE** (Taxdisc tab).

48) By default, the **BEST** results window now lists variable names (numbers are still optional) and also gives a summary table of the best solutions for each number of explanatory variables. Results are output continuously rather than at the end – as in all routines now.

49) Some routines are faster: the compute-intensive SIMPROF divides computation over multi-core processors; taxonomic distinctness will sample from large trees; and PCA is now SVD based.

50) Other minor changes are aimed at improving usability, convenience or speed of analysis.

a) The **Fill** and **Value** (or **Pattern**) menu, used when creating factor entries, now fills a partly blank and highlighted column in which the only needed entries to type in are the first one at each change.

b) The facility to open a very large workspace with its branches 'rolled up' in the Explorer tree – rather than all worksheets displayed (as at the last Save operation) – will allow a much more rapid start to a session, with only the analysis lines required then being unrolled.

c) Other changes include an improved **Print** dialog and the ability to restore the 'factory defaults' from the **Options** menu. Added Notes (right-click on the Explorer tree) now have fuller editing operations, e.g. for font type and colour, and can include images – another new worksheet type. **Duplicate** will now allow the copy to be placed directly below the original in the Explorer tree.

d) And there are new routines not even mentioned yet! (e.g. see **Variability weighting**, Section 4, a new pre-treatment option, and **Expand**, Section 14, to fill out model matrices to match biotic ones).

D. Typographic conventions for this manual

Emphases & text symbols

Text in **bold** indicates the menu items that need to be selected,

> denotes cascading sub-menu items, tab choices, dialog boxes or sub-boxes,

• denotes a button entry in a dialog box (so-called 'radio buttons' – only one can be selected),

✓ indicates a tick in the specified box (so called 'check boxes' – either on or off),

text inside a cartouche is an instruction to select the suggested entry (e.g. filename, factor etc) or actually to type it in, and

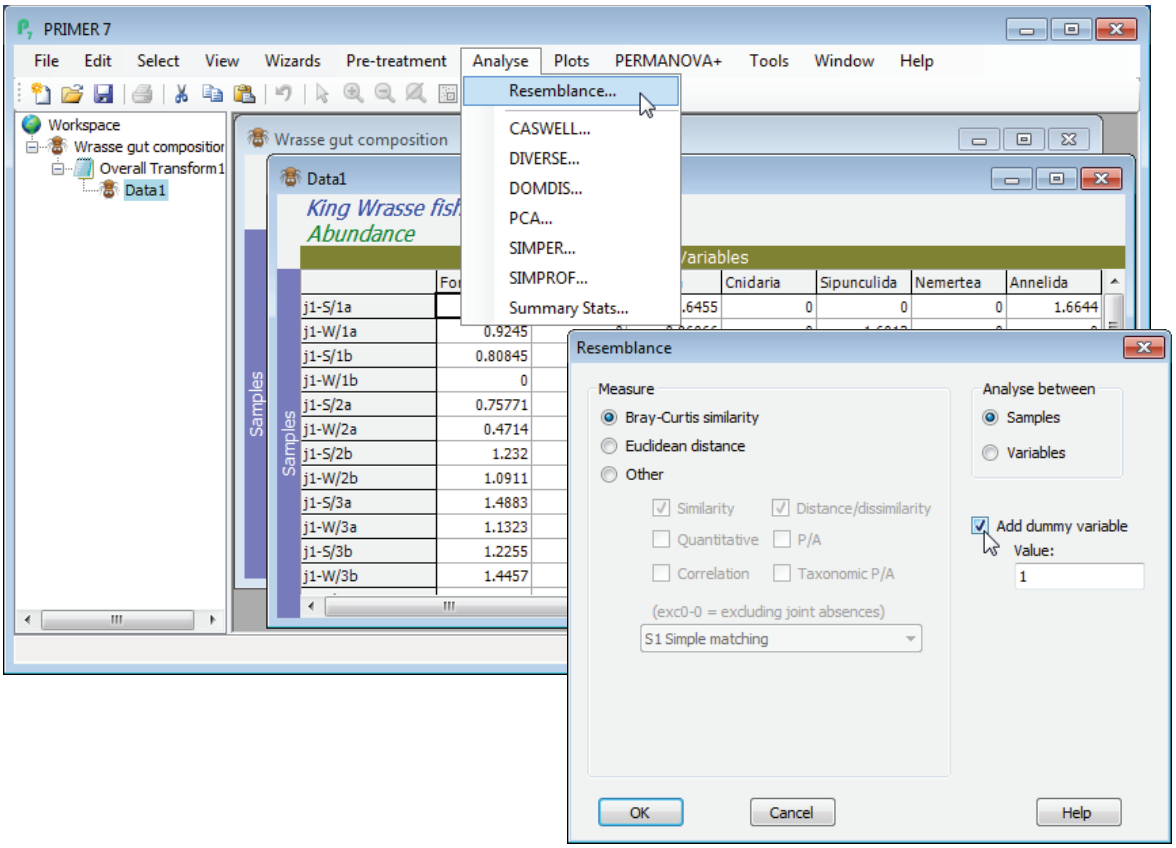
( ) & ( ) & ( ) indicate several steps that need to be carried out in the one box, where brackets are used naturally to split up the different components of the dialog.

In normal text, italics are generally reserved either for the first occurrence of technical terms, or for reference to factor names (when not in the context of entry to a dialog box), and underline is simply used for emphasis. Words in quotes have slightly transferred meanings, often only on first use (e.g. 'explanation' does not imply causality; 'replicate' implies use – but not collection – as a replicate).

For example:

Analyse>Resemblance>(Analyse between•Samples)&(Measure•Bray-Curtis similarity)&(✓Add dummy variable>Value: 1)

is an instruction to select the main menu item **Analyse**, the sub-menu item **Resemblance**, and to analyse between samples using Bray-Curtis similarity, adding a dummy species with value 1 for all samples, prior to computing similarities (this is the zero-adjusted Bray-Curtis, see Chapter 16 of the CiMC methods manual). The dialog this corresponds to is:



Finding your way around

Cartouches in the margin refer to the subsection headings listed at the start of this manual. It is also anticipated that navigation will be by searching on words or phrases in the PDF of this manual.

At the end of the manual there is an index of occurrences of each data set.

v7

A red vertical dotted line in the left margin indicates features which are new to PRIMER v7.

E. A brief tour through the operation of PRIMER v7

Opening the examples

After launching the PRIMER desktop by clicking on its icon, the first step is to open a worksheet of multivariate data, e.g. species abundances over a number of samples. The user's own data will typically be read into the program from Excel (*.xls or *.xlsx), though various text format input options are also provided (or you can type entries into a newly created PRIMER worksheet and edit it directly – though this is not commonly done). However, the PRIMER 7 installation comes with a number of real data sets, in a folder \Examples v7, needed as examples for this manual. You can access this with **Get Examples V7** on the **Help** menu, which prompts you for a directory in which to locate the \Examples v7 folder. It is assumed (for brevity) throughout this manual that this is simply the top level C:\ directory. So, C:\Examples v7 contains sub-directories for each study, in which the data files have usually been saved in PRIMER 7's internal binary format (*.pri, which is unreadable by other software or earlier versions of PRIMER). To open such a species × samples matrix, e.g. of nematode species abundances in marine sediments from 27 sites over five creeks of the Fal estuary, SW England (whose sediments are contaminated by heavy metals from historic mining), take **File>Open** from the main menu, navigate to the \Examples v7\Fal benthic fauna directory, select the file Fal nematode abundance.pri, and click **Open** to display the species data matrix in the desktop. Taking **Edit>Properties** you will see that PRIMER-format *.pri files carry other information on Title, Data type, Array size, whether Samples are found in •Columns or •Rows, and a Description. With **Edit>Factors** a subsidiary sheet of three factors is also seen to be linked to this worksheet: *Creek*, a creek abbreviation, the full *Creek name* and a numeric *Position* factor of the sampling sites' location down the creek – other factors could be typed in with **Add**.

Reading data in from Excel

As an example of reading in data from Excel, first open and examine the file Fal environment.xls, to note the simple format of a title in box A1, column headings (unique) for the samples in row 2, row headings (also unique) for the variables in column A, with only numeric entries in the array itself (rows 3 to 14). This is followed by a blank row, followed by the same three factors as above. This format must be adhered to precisely, with no extra blank rows or columns, or extra headers. After **File>Open**, you need to find the drop-down list (bottom right of the Open dialog) and select Excel Files, which should display the Fal environment.xls sheet. Select it, and **Open** now takes you through a File Wizard, for which you take the defaults – but look what they are – other than specifying (Data type•Environmental). If this worksheet were now to be saved in the default format (**File>Save Data As**), the result would be the Fal environment.pri file already in the workspace.

Basic MVA wizard

To cater for users completely unfamiliar with the basic outputs from a multivariate analysis, e.g. of the species abundance matrix opened above, PRIMER 7 now has a **Wizards>Basic multivariate analysis** menu item, which automatically generates robust outputs from some core routines, using knowledge of the Data type and with the opportunity for the user to alter some inputs from their defaults. Run this routine with the Fal nematode abundance sheet as the active matrix (click on it to make it active – its header bar will then be a slightly darker colour than other open worksheets). Take all the defaults on the Basic analysis wizard dialog box, i.e. just click on **Finish** – having first looked closely at the choices it has made for you! – and several results and graphic windows will appear in the display area of the PRIMER desktop (to the right). The Explorer tree area, to the left, shows the sequence of Data worksheets and Graph outputs created by the Wizard, interspersed with Results windows (the notebook icon) with names which describe the routine that has been run, and the tree shows the relationships among these analyses (what they start from and what they produce) A Wizard is just a bundled version of single routines which appear on PRIMER's other menus or sub-menus, so click on each row of the Explorer tree to display the sequence of steps involved and outputs produced. The final graphical step is a *Multiplot* output of four graphs, 'rolled up' in the tree, shown by the + sign. Clicking on the + (or on any of the plots in the multiplot) unrolls these names. It is now instructive to run through the individual analyses that this Wizard corresponds to.

Pre-treatment of data

Pre-treatment of the data (sometimes in more than one way) is usually desirable. For assemblage data, transformations will reduce the dominant contribution of abundant species to Bray-Curtis similarities. Though not usually needed for controlled ('quantitative') sampling, standardising of samples to relative composition (so sample totals are all 100%) can be achieved, with Fal nematode abundance active, by **Pre-treatment>Standardise>**(Standardise•Samples) & (By•Total) – the Wizard default was <u>not</u> to standardise but it did give that option, where % composition is desired.

Transformation of all values (which should be after standardisation, if the latter is appropriate) is obtained by, for example, **Pre-treatment>Transform (overall)**>(Transformation: Square root). A more severe transform would have been by Fourth root or Log(X+1) or by the ultimate in severity of transformation – reduction of the quantitative data to purely Presence/absence of each species. Since the purpose of transforming is to avoid the ensuing analysis becoming dominated by just one or two species with very large abundances, and bring more species into the definition of similarity of two assemblages – whilst at the same time avoiding giving sporadic, singleton species too much weight – the effects of competing choices can be assessed by running the second **Wizards** item.

Matrix display wizard

On active sheet Fal nematode abundance, run **Wizards>Matrix display**, not taking all the defaults in this case but unticking/unchecking the (Reduce species set) box so that all species are retained, and taking (Transformation: Square root) & (✓Retain sample groups>By Factor: Creek). A quite complex set of steps are then carried out, culminating in a run of **Shade Plot** from the **Plots** menu (fully described in Section 10) but, for our current purposes, all that needs to be understood is that the resulting shade plot is simply an image of the data matrix, in which the abundance for each species is represented by the shade of grey, from white (absent) to black (the largest count in the worksheet). Replicates from the 5 creeks are kept together along the *x* axis and the species on the *y* axis have been clustered and ordered in such a way that species with similar distribution across these samples are placed together in the re-ordering. (Multivariate analysis does not use the order of species in the matrix but it helps the human eye to visualise data structures by performing such re-arrangements). Apart from it being clear that some creeks contain a rather different set of species – or at least different abundances of the same species – an observation which is formally tested by the **ANOSIM** routine, e.g. as part of the **Basic multivariate analysis** wizard, the other message is that no one species will dominate an assessment of similarity of samples (columns) to each other. Equally clearly, quite a number of the less frequently occurring species have (transformed) values which are still sufficiently small in relation to the main players that they are almost invisible to the ensuing similarity calculation. This is probably desirable, and suggests we may have a reasonable transformation here. Contrast this with **Wizards>Matrix display** run again on the Fal nematode abundance sheet, but this time with (Transformation: None) – you can ignore the warning that PRIMER gives you (it is trying to tell you this is a bad idea!) – and it is clear from the resulting shade plot that only a few species will now contribute to the similarity computations. So an assessment of biotic differences among creeks, and how this relates to differences in heavy metal levels will really only be about a few numerically dominant species and not broadly community-based. At the other extreme, if you try the severest pres/abs transform, the rare species are now having far too much of an effect and will dilute genuine patterns from species sampled in reasonable numbers.

Section 4 also discusses an alternative approach to balancing contributions from different species, that of **Pre-treatment>Dispersion Weighting**, which downweights species with highly variable counts in replicates, which the sampling device captures in clumps rather than single individuals – relatively more weight is therefore given to species with consistent numbers over replicates of the same condition and these will be more reliable for assessment. If you try that pre-treatment and put the resulting rebalanced matrix into the **Matrix display** wizard, the shade plot gives a matrix image not unlike that for the square root transform, and this is certainly a possible pre-treatment here.

Environmental data

For environmental-type data, such as the Fal environment sheet, it is often appropriate to transform individual variables selectively, rather than all in the same way, since they may be of very disparate types. Here, the main objective is to avoid strong skewness in the distribution over samples, since large outliers will dominate both computation of (normalised) Euclidean distances and the Principal Component Analysis (**Analyse>PCA**), which is often the multivariate analyses chosen for abiotic data. The degree of skewness, or presence of outliers, is visually assessed using **Plots>Histogram Plot** or **Plots>Draftsman Plot** on active sheet Fal environment (you may wish to increase symbol size on the draftsman plot – do this by **Graph>Sample Labels & Symbols** and Size: 150, say). If there is strong right-skewness, those variables might need a log transform by highlighting them and taking **Pre-treatment>Transform (individual)**>(Expression: log(V+1)), Section 4. Alternatively take the rank transform, **Tools>Rank Variables**, which certainly gets rid of outliers! Although there is skewness here, there are no strong outliers and, for this demo, omit any transformation. So, run **Wizards>Basic multivariate analysis** on Fal environment and take all the defaults, examining the different choices made for this environmental-type matrix (e.g. normalising variables onto a common dimensionless scale; Euclidean distance resemblance; PCA ordination, see Section 12).

Resemblance calculation

Resemblance is the general term in PRIMER used to cover (dis)similarity or distance coefficients. The next stage in both the Fal nematode and environment runs of the **Basic multivariate analysis** wizard was to create an appropriate triangular resemblance matrix between all pairs of samples. This is a run of **Analyse>Resemblance** on the pre-treated (transformed or normalised) worksheet. Relevant defaults will be suggested, given the Data type, i.e. (Measure•Bray-Curtis) for biota and (Measure•Euclidean distance) for environmental variables, and (Analyse between•Samples) in both cases. There are, however, nearly 50 other possible choices on this dialog, see Section 5.

ANOSIM tests

The wizard then runs, for both biotic and abiotic data, **Analyse>ANOSIM>**(Model: One-way - A) & (Factors A: Creek)>(Type Unordered) on the respective resemblance matrices as active sheets. This tests for statistically significant differences overall among the 5 creeks in terms of their biota (or environmental data), and follows it up with pairwise tests between pairs of creeks, using the 5 (or in one case 7) locations in each creek as the replicate level. The Results window (eg *ANOSIM1)* shows the ANOSIM R statistic is large (0.82 for biota, 0.71 for environmental variables), close to its maximum value of 1, implying very good clear separation of the creeks, and highly significantly different from the null hypothesis R = 0, of no creek differences – the same is true of the pairwise tests. The associated plot (*Graph1*) is of the null hypothesis values of R under random permutations and shows that values not much more than R = 0.2 would be expected here if creeks did not differ.

ANOSIM tests can be much more extensive. PRIMER 7 introduces the idea of ordered ANOSIM tests, in which a numerical factor can be defined for the groups *a priori* (perhaps testing for simple time trend, or spatial gradient of change). Two-way crossed or nested, and three-way crossed, nested, or mixed crossed and nested, designs can be defined, with any factor ordered or unordered and analyses are then often possible without replicates as well as with them – see Section 9.

CLUSTER analyses

The Basic MVA wizards then run a cluster analysis, again on the respective resemblance matrices. This component routine is **Analyse>Cluster>CLUSTER>**(Cluster mode•Group average), without taking the (✓SIMPROF test) option since the latter is the appropriate test (rather than ANOSIM) when an *a priori* group structure is not defined. That is, if we had chosen to ignore the structure of sites within 5 creeks and simply treated the 27 samples as just 27 Fal estuary locations, the primary thrust of the analysis would not have been the ANOSIM tests and MDS display (see below) of those creek groups. Instead, it would have been a more exploratory analysis of whether the sites fell into clusters of similar communities (or environmental variables) at all – and, if so, which sites constituted those groups. The SIMPROF test is then important in deciding which sub-clusters in the hierarchical group-average cluster analysis (UPGMA) we are entitled to interpret as distinguishable groups, statistically – and, if we did not tick the ✓ANOSIM (1-way) box in the **Basic multivariate analysis** wizard, it would instead run a series of SIMPROF tests on the nodes of the cluster analysis dendrogram (Section 6) to determine this. As it is, the clustering in this Fal example is secondary and *Graph2* simply displays the dendrogram of the 27 sites, without SIMPROF tests. However, it is interesting to note that the dendrogram does largely divide the 27 samples into the 5 creeks, with an exception or two, which is consistent with the clear distinction among creeks seen in ANOSIM. You might like to accentuate this point by **Graph>Sample Labels & Symbols>**(Symbols✓Plot)> (✓By factor Creek) and look also at **Graph>Special** options, e.g. re-orienting the dendrogram.

PRIMER has other clustering tools (Section 6): a hierarchical binary divisive cluster analysis in unconstrained, **Analyse>Cluster>UNCTREE**, or constrained form, **>LINKTREE** (in which only divisions which have an 'explanation' in terms of a threshold on an environmental variable, say, are permitted). Both these methods share a common structure, consistent with the non-parametric treatment of resemblance matrices (which applies to tests such as ANOSIM, RELATE, BEST and ordinations such as non-metric MDS etc), namely each group is successively sub-divided so as to maximise the ANOSIM R statistic (PRIMER's key measure of group separation in multivariate space) between the two groups formed. A further non-hierarchical clustering method is available in the **Analyse>Cluster>kRCLUSTER** routine, a generalisation of classical *k-means* clustering to any resemblance matrix but again using only ranks. SIMPROF tests can be applied to all methods.

MDS & PCA ordinations

The Basic MVA wizard next produces non-metric MDS (*n*MDS) plots in 2-d and 3-d, together with their associated Shepard diagrams, which show how well (or badly) these distances among samples in the low-d ordination plots approximate the high-d resemblances. If the stress (Section 8) is not too large (it is only 0.10 here), *n*MDS plots give a powerful representation of the sample patterns.

The wizard is here running (again on the resemblance matrix) **Analyse>MDS>Non-metric MDS (nMDS)** under default conditions, but taking this directly, there are options to choose higher-d solutions and, more entertainingly, to watch the iterative process of trying to obtain the lowest stress 2-d solution (say) from different restarts of sample points thrown randomly into 2-d, which you can activate – and even record as an *.mp4 file – by (✓Animate) on the dialog (Section 8). There are other recordable animations possible also, of spinning 3-d ordination plots and showing a dynamic trajectory of, for example, a time series of samples on an MDS or other ordination plot.

Where the data matrix is environmental and (usually) variables normalised, there is a choice of ordination by PCA (**Analyse>PCA** run on the normalised data sheet) or *n*MDS on the Euclidean distance resemblances. These are both offered by the wizard but, running directly, a third option is **Analyse>MDS>Metric MDS (mMDS)**, which fits a straight line to the Shepard diagram of MDS (low-d) distances *vs* original (high-d) distances, and in one respect improves on PCA here, giving a more faithful preservation of the high-d distances by avoiding the PCA projection into low-d.

Species analyses

The final step in the Basic MVA wizard is to break down the dissimilarities (or distances) between pairs of creeks into their contributions from each of the species (or abiotic variables), in the tables of *SIMPER1* (or *SIMPER2)*, see at the end of Section 10. This is equivalent to running **Analyse> SIMPER** on the transformed data matrix for biota (or normalised data matrix for abiotic variables). There are, however, several other ways in which PRIMER examines variable relationships to each other, or species relationships to the sample patterns (Section 10). We have already seen the power shade plots potentially have for interpretation. Another possibility is Bubble plots of individual species values on the sample *n*MDS ordination: the larger the bubble the greater the abundance of that species at that site – or abundances, because PRIMER does multiple (segmented) bubbles of different colours and circle sectors for different species. For the Fal *n*MDS, try this with **Graph> Special>**(✓Bubble plot) & (Worksheet: Fal nematode abundance) & (Variables>**Change**), moving *Metachromadora vivipara*, *Tripyloides gracilis* and *Leptolaimus limicolus* to the Include box and all other species in the Available box, and ticking (✓3D effect) & (Saturation: 75).

Calculating similarities (index of association) among species – not samples – or correlations among environmental variables, in their pattern of response across the samples, opens up another field of analyses, which we have already seen used to cluster species in the shade plot. Adapting SIMPROF tests to operate on variable clusters (Type 3) rather than sample clusters (Type 1) permits definition of *coherent variable sets*, which within the sets are not statistically distinguishable but across sets have significantly different response patterns over the samples. Run **Wizards>Coherence plots** on Fal environment, the heavy metal levels (and silt/clay ratio) at these 27 sites, with significance set at 0.5%, and strikingly similar metal concentration profiles are seen in the resulting **Line Plot** sets.

Other analyses

A further bubble plot you might like to try on the Fal *n*MDS is to superimpose abiotic variables from the Fal environment worksheet, and we have already referred to the constrained LINKTREE clustering that tries to explain community groupings in terms of particular environmental variables, but PRIMER also has another generic way of looking at the relation of community structure to potentially explanatory variables – in combination, rather than individually, see **BEST** in Section 13. There is an overall hypothesis test for the significance of such a link, and the mechanism of non-parametric matrix correlations (which also includes PRIMER's **RELATE** tests) can be applied to other contexts in which multivariate data sets are compared (Section 14).

PRIMER calculates a range of univariate diversity-related indices through the **DIVERSE** menu including ones based on taxonomic or genetic/functional relatedness of the taxa (**TAXDTEST**), see Section 15, and a range of diversity curves (eg dominance plots, species accumulation, Section 16).

Other main menus (e.g. **Select**, **Edit**, **Tools**, **Plots**) offer a wide variety of data manipulations and standard plotting functions (**Histogram**, **Means**, **Box**, **Bar**, **Surface**, **Line** and **Scatter Plot**).

The final Section (17) deals with region estimates for means in multivariate studies, e.g. average communities for each of the Fal creeks, plotted on a 2- or 3-d MDS together with an approximate measure of the uncertainty about these means, from bootstrapping. You might like to finish this brief excursion through PRIMER by running **Analyse>Bootstrap Averages** on the Bray-Curtis resemblance matrix from the Fal biota, taking all the defaults, to get the multivariate *means plot*.

And you can save all this in a PRIMER workspace file *.pwk with **File>Save Workspace As**.

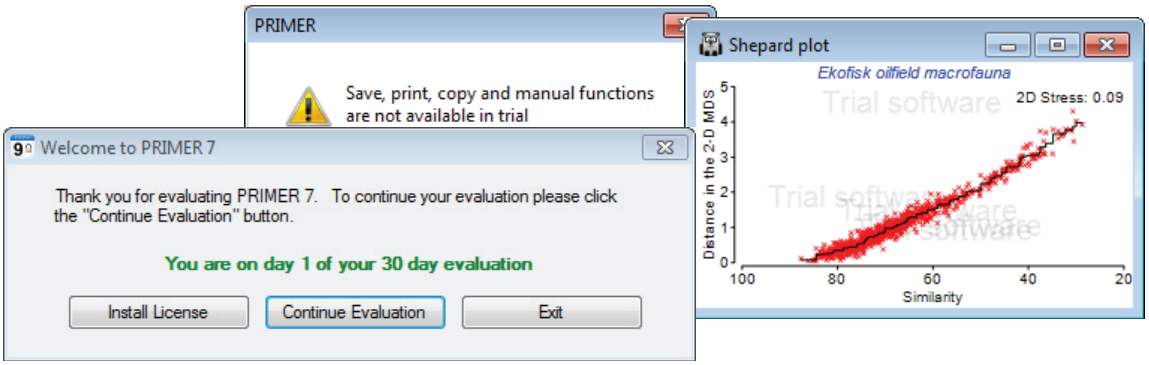
MANUAL/TUTORIAL

0. Trial version, Help system, Manuals, Updates, Install and Uninstall (*Help*)

PRIMER 7 trial software

v7

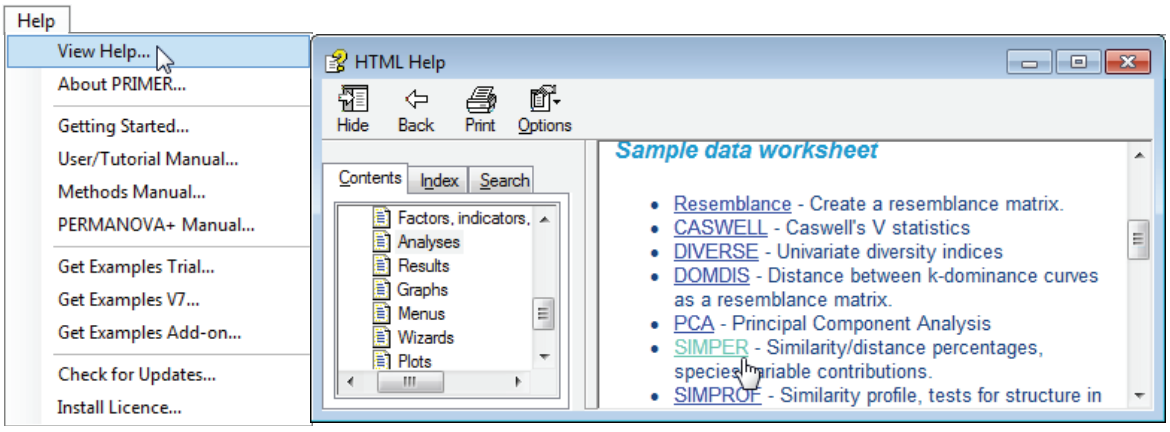
A trial version of PRIMER 7 is freely available, which is downloadable from the PRIMER-e web site (www.primer-e.com), and can be run without an installation key. Whilst it is a full working implementation of all analysis routines and graphics, and permits a user's own data to be input from Excel or text files and analysed, it is only intended to be trial software so there are a few key limitations – all printing, copying to the clipboard and saving of data and workspaces is disabled. In addition, plots are watermarked so presentation-standard output is unobtainable from screen grabs. The trial version has a 30+ day expiry date, with an additional grace period, but will then cease to operate – reinstallation of the trial version on the same machine is not possible.



A valid installation key can be purchased at any time from PRIMER-e and inputting that to the Install routine when on-line (see below) will remove all such printing and saving restrictions and watermarks, and any updates and download of the PDF files of the manuals will then take place. The trial software includes both PRIMER and PERMANOVA+ but, if only a key for PRIMER is purchased, PERMANOVA+ routines are not enabled and their menu will no longer appear. If a single-user PERMANOVA+ licence, for use with PRIMER 6, is registered to you then no further purchase of PERMANOVA+ is needed – it operates in is essentially the same way with PRIMER 7 as it did with PRIMER 6, and you will be given a single installation key enabling both products.

Help system & manuals

The **Help** main menu has the entries shown, starting with **View Help**, the HTML Help system:



This lists and describes all menu items but in succinct fashion since (in addition to a brief **Getting Started** manual which comes with the trial version) there are two comprehensive pdf manuals for PRIMER, and one for PERMANOVA+, opened (if enabled) by the **Help** menu items:

User/Tutorial Manual (this manual): Clarke KR, Gorley RN (2015) *PRIMER v7: User Manual/ Tutorial*, PRIMER-E, Plymouth, 296pp;

Methods Manual: Clarke KR, Gorley RN, Somerfield PJ, Warwick RM (2014) *Change in marine communities: an approach to statistical analysis and interpretation, 3rd edition*, PRIMER-E, Plymouth, 260pp;

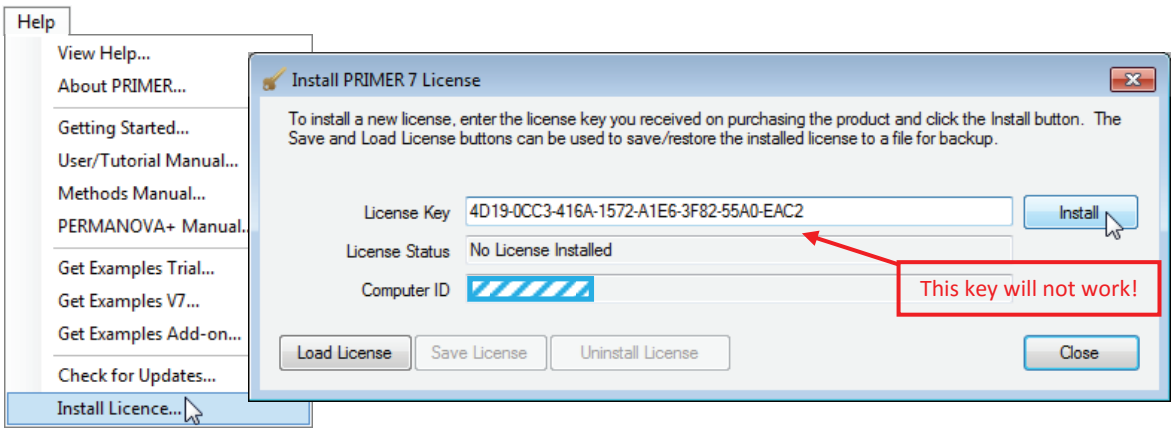
PERMANOVA+ Manual: Anderson MJ, Gorley RN, Clarke KR (2008) *PERMANOVA+ for PRIMER: Guide to software and statistical methods*, PRIMER-E, Plymouth, 214pp.

Updates

The **Help>About PRIMER** menu item will tell you: which maintenance version number you are currently operating of PRIMER 7; the licence type, e.g. Trial/Full/Academic; and (if it is not a trial version) the unique serial number registered to you in the PRIMER-e database. The **Help>Check for Updates** menu item tells you whether you are running the latest maintenance version, i.e. if an update is available. Unless you have changed the default in **Tools>Options>Updates** away from (✓Automatically check for updates) you will already have been prompted to update, on release of that maintenance version. Typically, an update will not usually take much in excess of 3 minutes. You will be required to agree to the (standard) licence terms and conditions.

Install and Uninstall

You will need to run **Help>Install Licence** at the point at which you have purchased an installation key (32 characters in blocks of 4), which is copied and pasted into the License Key box seen below. Once you have installed this key you do not need it again in order to download an update – the two operations of downloading/updating the software and authenticating your PC to run it (the latter being what the installation key is all about) are quite separate processes.



You might also need the above dialog to **Uninstall Licence** (i.e. de-authenticate it), e.g. if you get a new PC and need to move the software to it from your old one. All PRIMER 7 single-user licences will allow two simultaneous installations. To transfer an authentication to a new PC, on the old PC you must either run **Help>Install Licence>Uninstall Licence** (the inoperable software is then left on the machine) or take the standard Windows *Uninstall a program*, removing the software as well as the authentication. There is, however, a limit to the number of Uninstall/Install cycles permitted.

The primary need for **Save License** and **Load License** is in off-line authentication (for which you need to contact the PRIMER-e office). Note that they do not offer a means of saving the key and using it to authenticate the software on a new machine – you need the full License Key for that. Your main flexibility to cope with problems (e.g. a serious unrecoverable machine failure which stops you carrying out an Uninstall) is the second simultaneous installation permitted for the key. Or you may prefer to use the second licence for a home machine or a travel laptop, but keep your installation key secure (give it to someone else and you lose control of your installation options!)

Example data

Help>Get Examples V7 will take you to the data sets used throughout this manual [and **Help>Get Examples Add-on** will make available the data sets used in the PERMANOVA+ manual]. You are prompted for a directory in which to store the data files, i.e. a location in which you will find it convenient to access them in future, under the folder \Examples v7 [or \Examples add-on].

1. Opening, editing and saving data (*File*, *Edit*)

Getting the examples

The installation and subsequent run of **Help>Get Examples V7** will have placed a number of sub-directories (BC zooplankton, Bermuda benthos, …, Wrasse diets) into the \Examples v7 directory, placed in a location which you have chosen. Throughout this manual it is assumed (for brevity) that the Examples v7 directory has been placed at the top level, i.e. the folder is C:\Examples v7. The various sub-directories contain the faunal matrices and, for some sets, the matching environmental data and taxonomy, for most of the case studies described in the Methods manual (CiMC: 'Change in Marine Communities', 3rd edition, 2014), and all the data sets in this manual. Check this by making sure you can find the directory \Examples v7\Fal benthic fauna, of the soft-sediment core samples of biota and matching environmental data for 27 sites from 5 creeks of the Fal estuary, subject to different levels of heavy metal contamination. That directory contains files such as Fal copepod counts.pri and Fal copepod taxonomy.agg, with similar files also for macrofauna and for the meiofaunal nematodes, all in internal binary PRIMER 7 format, and Fal environment.xls, an Excel sheet of variables × samples (though this could equally well be of samples × variables).

PRIMER file types

Whether the extensions *.pri, *.xls(x) etc display, or not, is a function of your Windows set-up; if you have suppressed them it is still easy to distinguish the different file types by their icon. There are PRIMER 7-specific icons and extensions for the following:

 (*.pri) sample data in rectangular format, and associated factors, description etc;

 (*.sid) triangular matrices of similarity, dissimilarity, distance (generically, resemblances);

 (*.agg) aggregation file, assigning species to genera to families etc;

 (*.ppl) plot files, holding all the internal PRIMER information that structures the plot;

 (*.pwk) workspace files of everything, so the PRIMER desktop can be reconstructed 'as was';

[ (*.pdd) design files – these are only used in the separate PERMANOVA+ routines which are an add-on to the PRIMER software, see the separate PERMANOVA+ User Manual].

Compatibility of files

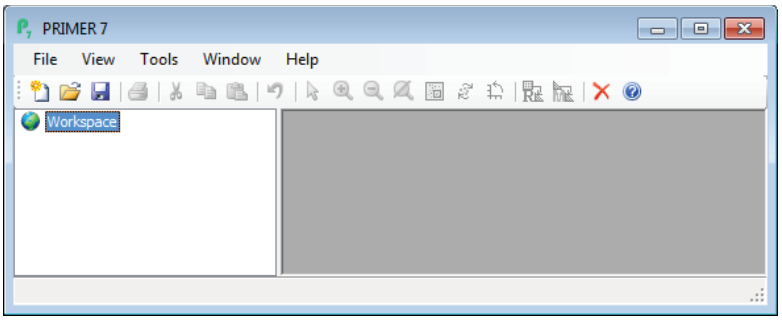
PRIMER 7 is fully forward-compatible from v6, and can input PRIMER 6, 5 and 4 data (*.pri and *.pm1), similarity (*.sid and *.sim) and aggregation (*.agg and *.pm1) files, and PRIMER 6 plot (*.ppl) and workspace (*.pwk) files, directly. But it is not, in general, backward-compatible so that PRIMER 6 cannot read workspaces, plot or data files saved in v7 format – there are many new plotting routines in PRIMER 7 that require additional internal data and graphic storage structures, workspaces from v7 in the earlier v6 format; little is lost with *.pri data files but, when such workspaces are re-opened in v6 (or v7), plot formats which are new to v7 will not be present, clearly. In addition, PRIMER 7 recognises several standard Windows extensions, e.g. input and output of *.xls(x) (Excel) and *.txt (text) data files or resemblance matrices; also *.csv, comma-separated text input. There are several options for *.txt text-format data input. Results windows can be saved in *.txt or *.rtf (rich-text) formats. PRIMER 7 can output plots to standard bitmap formats (*.bmp, *.jpg, *.png, *.tif, *.gif). An image in one of these formats can also be input to v7 (though with only minor use, as an image in an MDS bubble plot). Most usefully, plots can be output as vector-based enhanced metafiles (*.emf), and Copy and Paste operations from Graph windows in PRIMER to Microsoft Office routines (e.g. Powerpoint) are of this type, so plots can be ungrouped into Office drawing objects and further manipulated. PRIMER 7 also introduces some animation options, e.g. for spinning 3-d ordination plots, which can be output as movie files (*.mp4) or animated GIF (*.gif). All other file types (extensions) are not recognised.

v7

v7

Opening the PRIMER 7 desktop

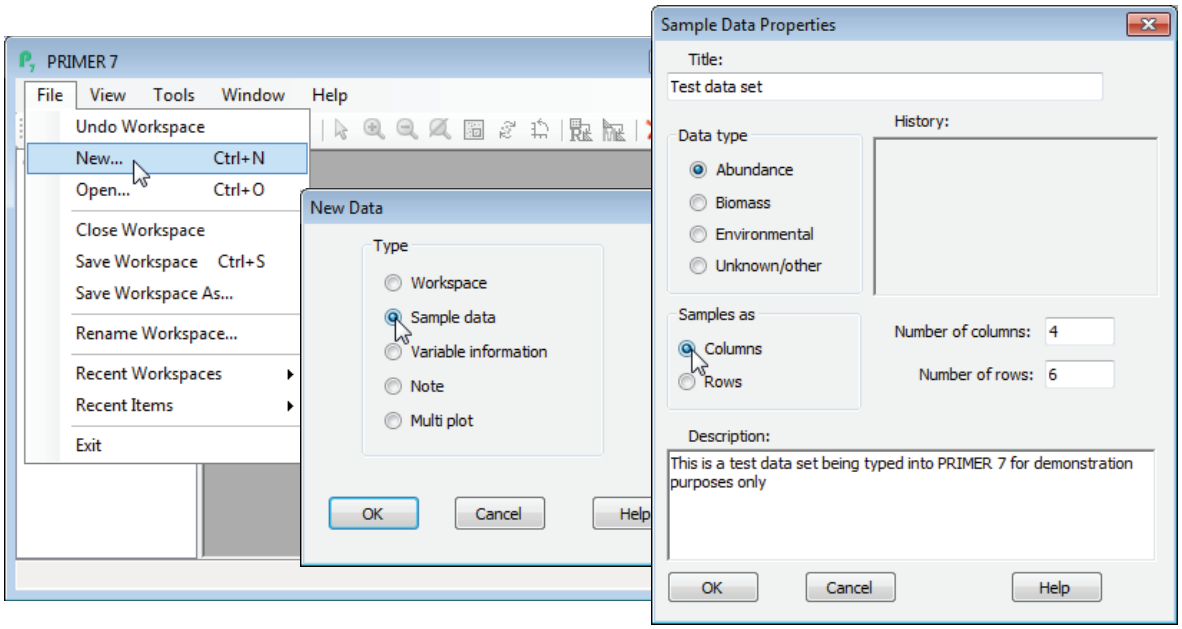
Start the program by (double)-clicking on the desktop or task bar PRIMER 7 icon  , giving the window below. A second method is to double-click on a file with a recognised PRIMER extension, e.g. a worksheet file (.pri) or a workspace file (.pwk), and PRIMER 7 will automatically launch, with the selected file or workspace placed in the resulting desktop window. Note that opening more than one sheet by Windows Explorer>**Open** on a selection of filenames launches parallel PRIMER desktops, which is usually not the required outcome. To open multiple files simultaneously into the same workspace, first launch PRIMER then select several files in the **File>Open** dialog window.



The PRIMER desktop is separated into two parts: to the left is the Explorer tree which will display icons for all the sheets, results windows, plots etc, and their interconnections. To the right, the actual worksheets, results and plot windows are displayed. The current workspace consists of all items in the Explorer tree (irrespective of which windows are displayed on the right hand side of the desktop), and all files needed for an analysis must first be opened into the current workspace.

Entering data directly

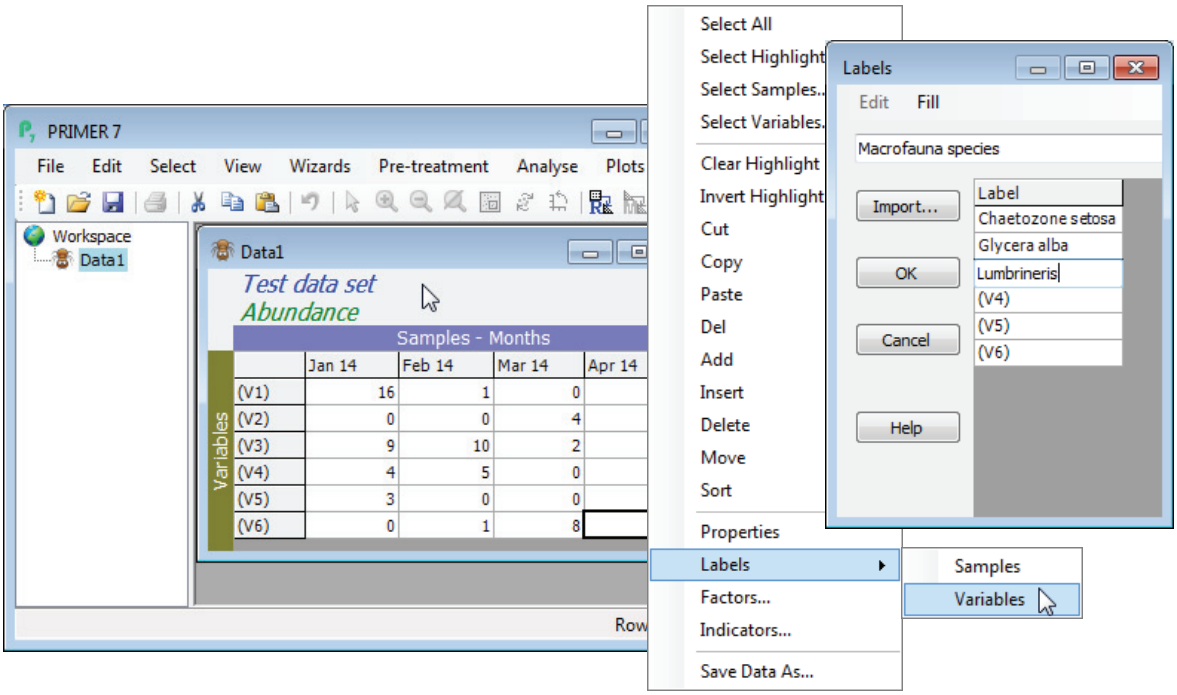
Most users will already have their data stored in rectangular form in some other software, e.g. as an Excel spreadsheet, which can be opened directly and straightforwardly (see later). However, data arrays can be typed directly into PRIMER, if necessary. Select **File>New>**(•Sample data) and, in the resulting Sample Data Properties dialog box, type in a title, specify a data type and which way round the matrix is to be, e.g. (Data type•Abundance) & (Samples as•Columns). You can also give a description of the data (optional), and you need to enter the number of columns and rows.



A worksheet of zeros is created into which you can type, by working down the columns, clicking on the first cell, typing in the number and pressing the Enter key. To edit an earlier entry, double click on the cell, amend it and again press Enter. To cancel an edit of a cell you have entered by mistake, press the Esc key. If you inadvertently click on a row or column label (the grey cells at the margins of the table) that row or column will be highlighted; remove the highlighting by clicking again on the label (highlighting is a simple on-off 'toggle').

Labelling samples & variables

At this point only the default row (variables V1, V2, …) and column labels (samples S1, S2, …) have been defined, but a set of commonly used operations for worksheets can be found in the lower part of the **Edit** menu, including a **Labels** item. This menu can also be called up by right-clicking when the cursor is placed within the body of the worksheet (see below). The samples or variables can then be labelled: labels benefit from being succinctly descriptive; they must be spelt consistently from one sheet to another and unique within a sheet. This is because PRIMER 7 makes much use of label matching (e.g. abundance to biomass or species to environmental variables at the same set of sites, or the merging of species lists from different studies etc).



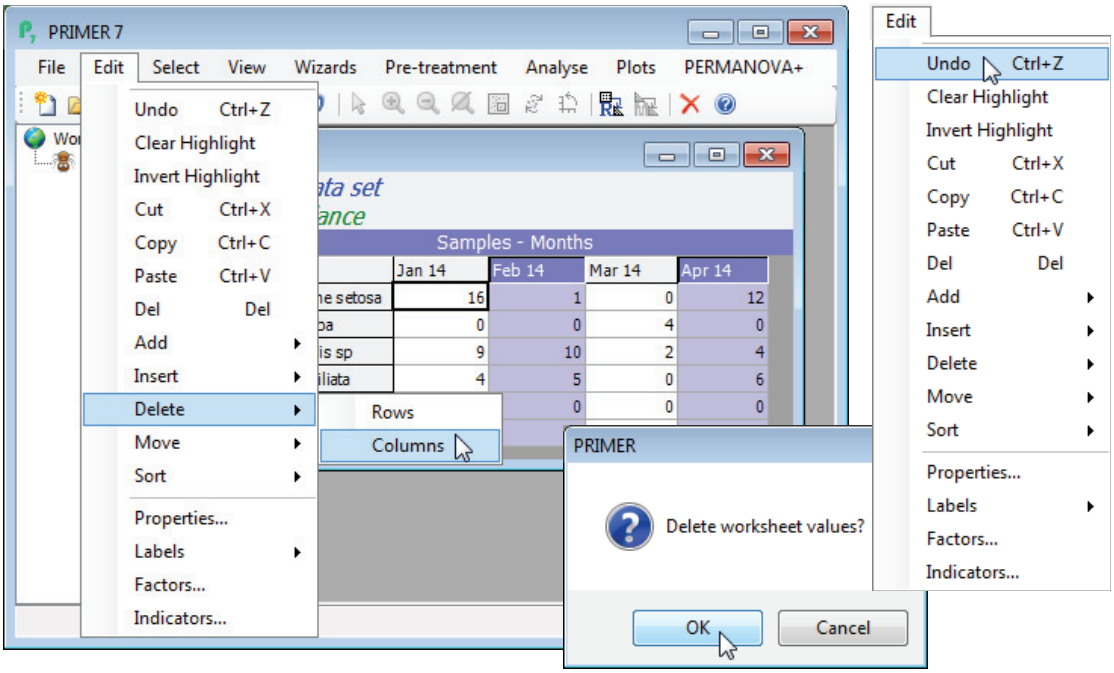
Deleting & inserting rows/cols

In addition to attaching labels, the **Edit** menu allows a range of other edit functions on the data entries. For example, to delete whole columns (or rows), highlight them by clicking on their labels and then take a menu sequence such as **Edit>Delete>Columns**. A prompt is given for all such deletion operations to ask whether they were really intended. Note that the current cursor position – the cell in the sheet outlined in black – is ignored; deletion works only on highlighted rows or columns. In contrast, insertion ignores highlights and uses only the current cursor position, e.g. **Edit>Insert>Row** will add a new row immediately above the current position of the cursor and **Edit>Insert>Column** adds a new column to the left of the cursor. Logically therefore, if a new row or column is needed at the bottom or right of the whole data sheet, respectively, a different operation is required: this is **Edit>Add>Row** (or **Column**) and it will ignore the position of the cursor or any highlighting of rows or columns.

Undo data sheet edits

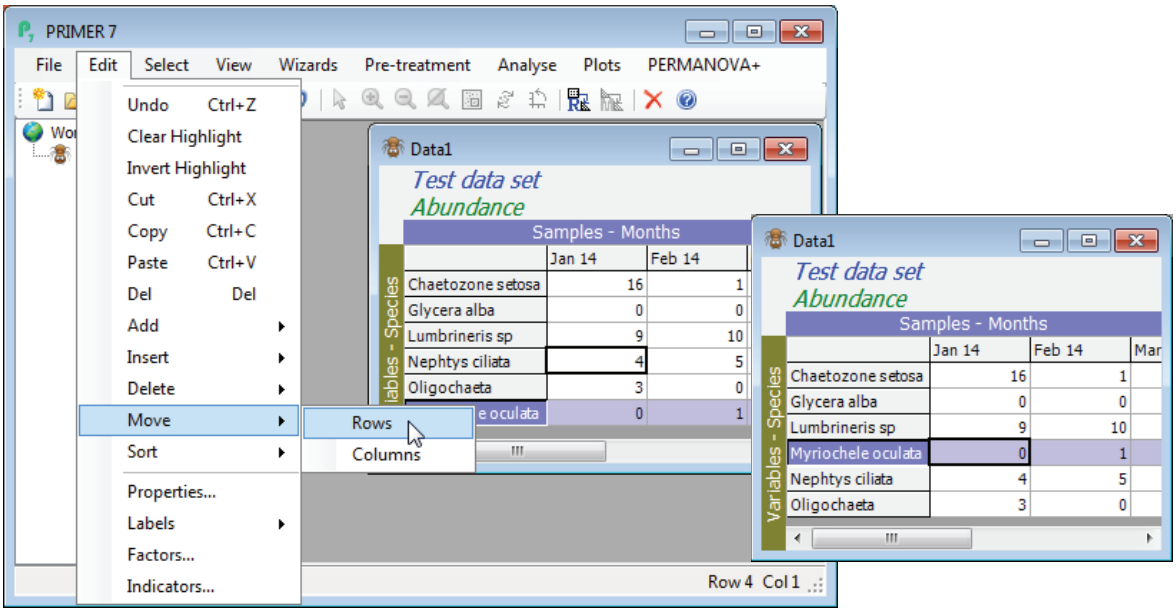
v7

Occasionally, cell values are accidentally deleted or rows/columns added in unintended places but PRIMER 7 (unlike PRIMER 6) is able easily to back-track on such changes. There is a repeated **Edit>Undo** operation for all row and column manipulations on the **Edit** menu (between the **Cut** and **Sort** items). The **Undo** extends to any typing directly into cells of the data sheet or to copying and pasting rows/columns into it, whether from the same sheet or externally, from the clipboard.



Moving & sorting rows/cols

Movement of rows or columns uses both highlighting and the cursor position. Rows (or columns) to be moved are highlighted, and the **Edit>Move>Rows** operation moves all highlighted rows to immediately above the current cursor position when moving up, and below the cursor position when moving down (similarly with moving columns to the right or left – movement is always over the cursor). In the simple case illustrated below, the same outcome would have been achieved by **Edit>Sort>Rows>**(•By labels), since this is an increasing alphabetic sort of the row labels. Note that sorting can also be carried out according to some alphanumeric order, held in an *indicator*. The latter is the term that PRIMER uses for information associated with each variable – a catalogue numbering system here perhaps – see Section 2 on setting up factors and indicators).



Cut, copying & pasting

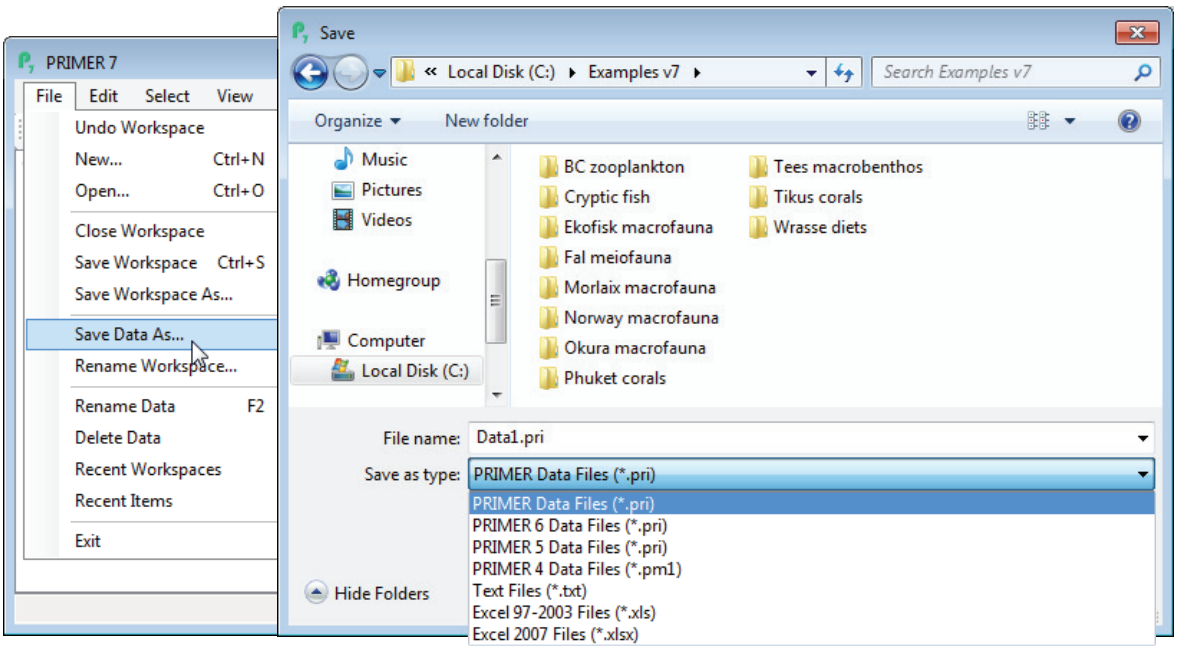
The **Edit>Cut** and **Edit>Copy** operations send part of a worksheet (or its factors/indicators) to the clipboard, where they are accessible by other Windows software or can be pasted back into another region of the active sheet (or factors). **Cut** and **Copy** operate much as in Excel etc, on highlighted regions of the worksheet, the difference here being that highlights must be whole sets of rows, or columns, or a combination of rows and columns, the highlighted data always being the darkest displayed cell colour (if nothing is highlighted the cell at the current cursor position is copied). **Edit>Paste** places the data from the clipboard onto a highlighted area of the same shape or, if there is nothing highlighted, onto a rectangle with its upper left corner at the current cursor position.

Saving data, renaming & deleting

The data sheet can now be saved (as can any item created in the workspace) from the **File** menu. **File>Save Data As** gives a standard type of Windows dialog box, shown below. This allows you to change to the desired directory, specify a meaningful name for the file (the default is Data*n*, where *n* just numbers each new data sheet in increasing order) and save it in PRIMER 7 format, with .pri extension, e.g. here in C:\Examples v7 and (File name: Test1) & (Save as type: PRIMER Data Files (*.pri)). This is the standard (binary) format for PRIMER 7 data matrices, which cannot be read by earlier PRIMER versions (including PRIMER 6) or by other software, but it is possible to choose to output in these earlier formats: PRIMER 6, 5 (Windows *.pri binary files) and 4 (the original DOS *.pm1 text file), plain text files (*.txt), and either the earlier Excel format (*.xls) with its restriction to 255 columns or the post-2007 *.xlsx format with no such constraints. On the same File menu there is an option to **File>Rename** the data sheet in the current workspace, or you can simply click once on the highlighted name box in the Explorer tree and overtype or edit the name. It is often a good idea to change the standard default names to something more meaningful in the context, so that you can find your way round the workspace more easily. Another option is **File> Delete**, which not only removes the specific data sheet from the workspace (though does not delete it in the original directory of course!) but also removes all the structure which leads immediately from that sheet (data sheets, results or graphs on the same branch of the Explorer tree, Section 7).

Undo in the workspace

Another new feature to PRIMER 7 is the option **Undo Workspace** which back-tracks for the work-space operations of renaming or deleting any sheet (or other window), or renaming the workspace.

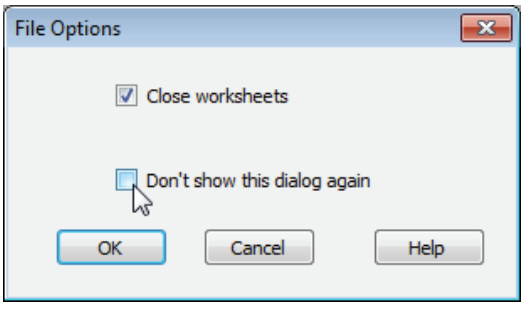


Saving, closing & opening a workspace

Typically, a single workspace would encompass one or more inter-connected data sets which are analysed at the conclusion of a specific phase of a study (unrelated data sets and analyses are best handled in separately created workspaces). Thus, rather than saving individual data or results files, more useful is the ability in PRIMER to save entire workspaces. This can be accomplished initially with **File>Save Workspace As**, which gives a Save dialog box similar to the above, but producing *.pwk file types, which are internal binary formats not accessible by any other software. By default this saves a PRIMER 7 workspace format, capturing all the data, results and graphical structures, and their links, as represented in the current state of the workspace. (Subsequent workspace saves with the same name, overwriting the existing copy, are carried out with **File>Save Workspace**).

v7 An alternative is to save the workspace in PRIMER 6 format (also with a .pwk extension), within which many of the new features in v7 cannot be represented and are therefore omitted, though the file can then, of course, be opened by users working only with the earlier PRIMER 6. In general, it is only the new graphic formats which are lost, with the basic tree structure of the analyses and the results windows being retained. More detail about managing workspaces, and exploring the links between component items, is given in section 7 but, for now, note that other functions frequently used are **File>Close Workspace** (or **File>New>Workspace**), which leaves a new, clear workspace ready for opening of new files (and will prompt for a Save Workspace operation unless one has just been carried out), and opening an already saved workspace (either of PRIMER 6 or 7 format) with **File>Open**. For the latter, after the usual dialog to select a *.pwk workspace file, taking the tick box option (✓Close worksheets), see below, suppresses the roll-out of the branches of the Explorer tree, so that just the top-level worksheets in the workspace are shown, prefixed by ⊞, and sheets are not displayed on the PRIMER desktop. (Clicking on the plus signs, successively, will unravel the full tree structure). This is a potentially useful new feature in v7 which speeds up opening of a large workspace, which was saved with very many sheets open on the desktop. It is not the 'factory default' but once selected this becomes the default option for opening the next workspace, even for a new run of PRIMER. For a consistently preferred option here, this dialog box can be eliminated.

v7

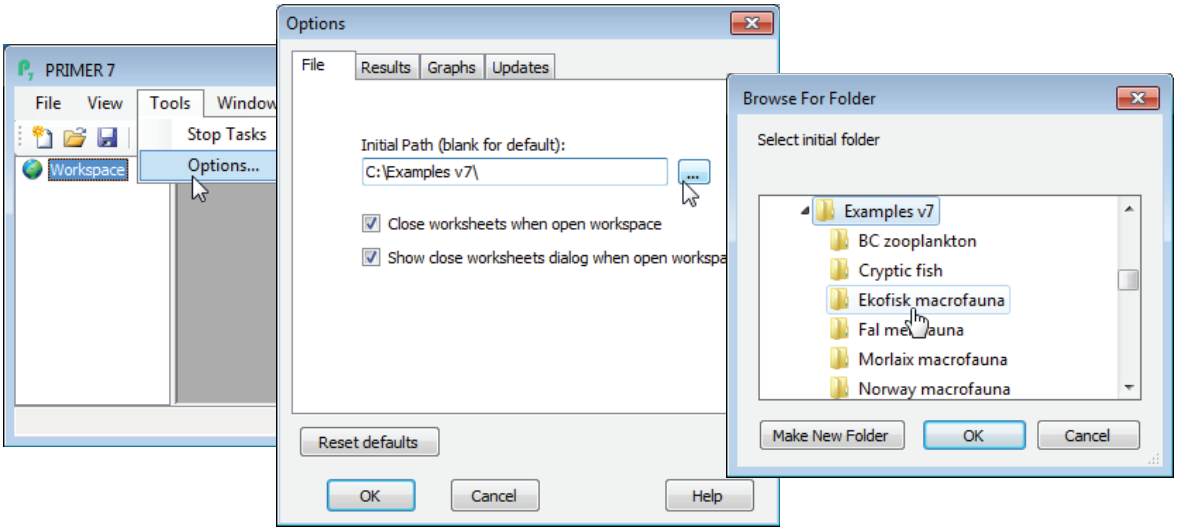


The test data matrix saved above is too small to do anything useful with, typical species by samples matrices often having hundreds of species or samples. So, take **File>Close Workspace** and we shall now work with a real 174 species by 39 samples array (Ekofisk macrofauna counts.pri), of soft-sediment macrofaunal abundances, in PRIMER 7 format, and a matching 39 samples by 9 environmental variables matrix (Ekofisk environment.xls), in Excel format, both files being in the directory C:\Examples v7\Ekofisk macrofauna.

Setting the initial directory

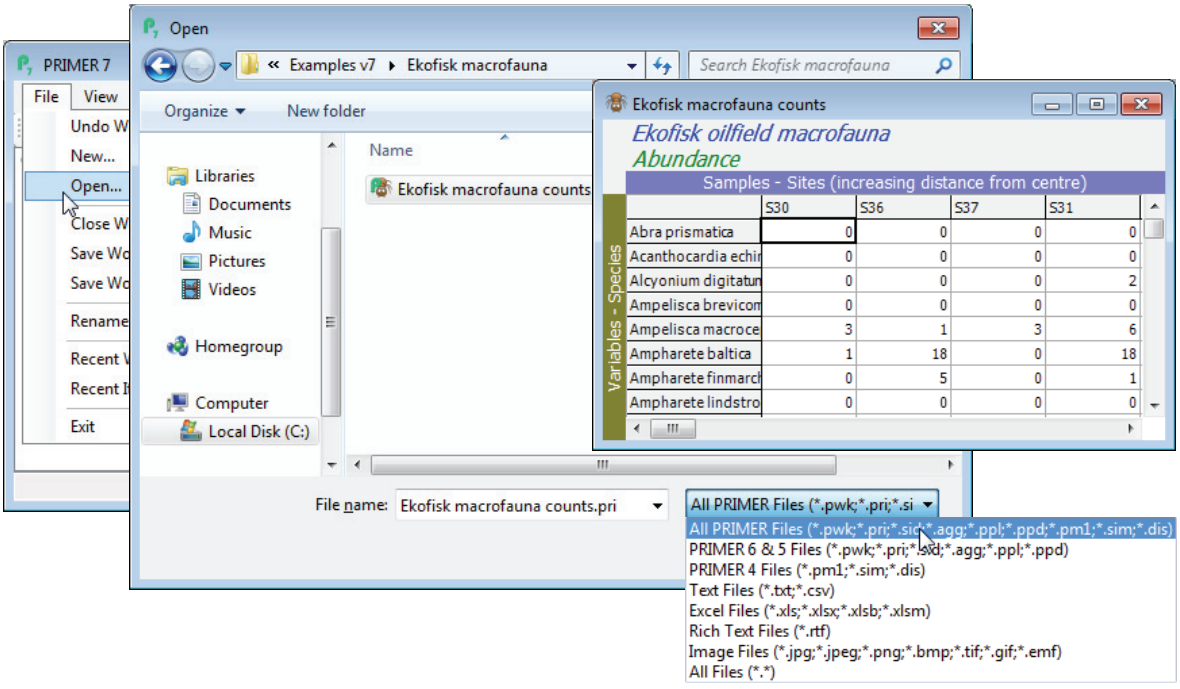
v7

It may sometimes be convenient to set the initial (default) directory to which PRIMER 7 opens every time it is run. Here this would use **Tools>Options** and supply (or browse for) the directory C:\Examples v7\Ekofisk macrofauna, see below. (Note that this run of PRIMER has to be shut down before the default change is implemented for future runs of PRIMER). For this tutorial, the default might logically be set to C:\Examples v7. Often, however, it is more convenient to leave this box blank and the program then always defaults to the previously used directory. (Incidentally, the **Tools>Options** dialog also gives the option to reverse the decision made in the illustration above, to eliminate the (✓Close worksheets) dialog box on opening a saved workspace.)



Opening PRIMER files

File>Open>(File name: Ekofisk macrofauna counts.pri)>Open will read in the existing Ekofisk species-by-samples array. The default file types are any of the PRIMER 7 formats, though earlier formats (Windows PRIMER v6 & v5, DOS v4), text and Excel files (commonly used for initial data input) are available. These are listed in detail on the first page of this section (Section 1).



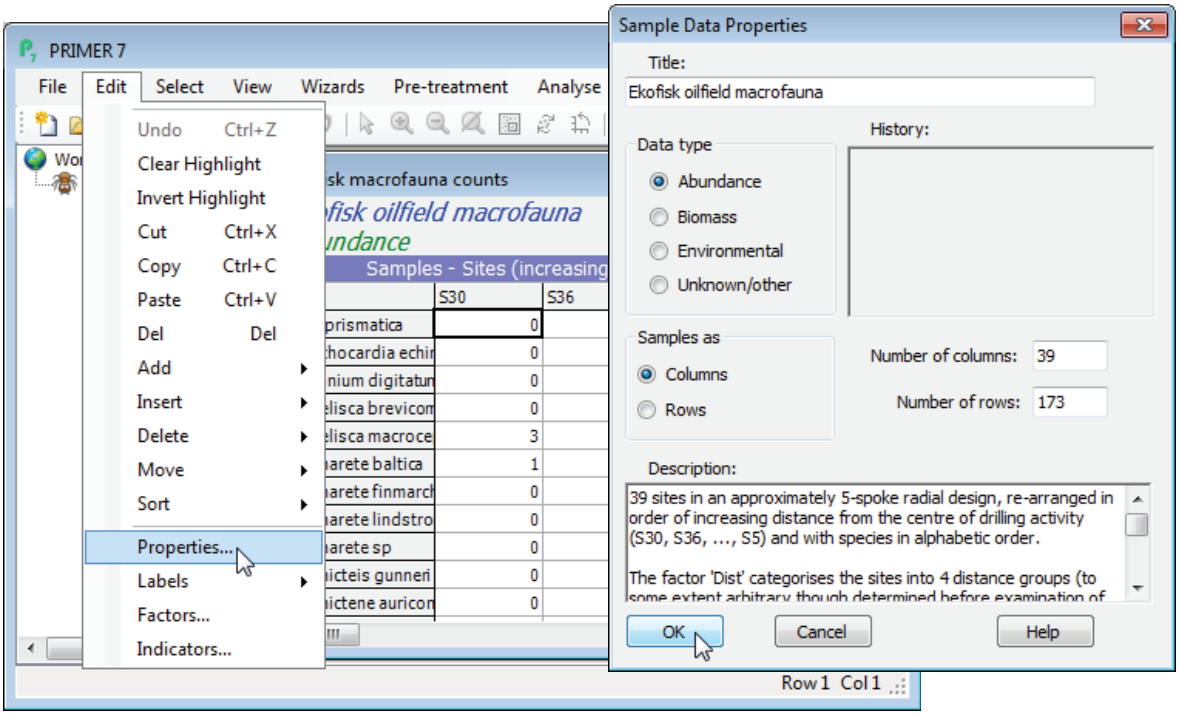
(Ekofisk oil-field fauna)

The abundance file Ekofisk macrofauna counts.pri is displayed within the PRIMER desktop. It shows the typical sparseness of species matrices, with many zeros and some large counts, each sample being the total of three Day grab samples at each site and the sites, S30, S36, S37, S31, … being ordered left to right in increasing distance from the centre of oil drilling activity – the design is roughly that of five radial transects from the centre of the oilfield out to distances of 8 km, at geometrically increased spacing. See Chapters 10 and 14 of the CiMC manual for a diagram of the sample sites and other details of this study (original paper: Gray JS, Clarke KR, Warwick RM, Hobbs G 1990, *Mar Ecol Prog Ser* 66: 285-299).

The data matrix, and any window in the PRIMER desktop, can be resized by dragging a corner of the window, as for the desktop itself, in normal Windows action. Note that by clicking on any cell in the matrix, its row and column numbers are indicated at the bottom right of the desktop.

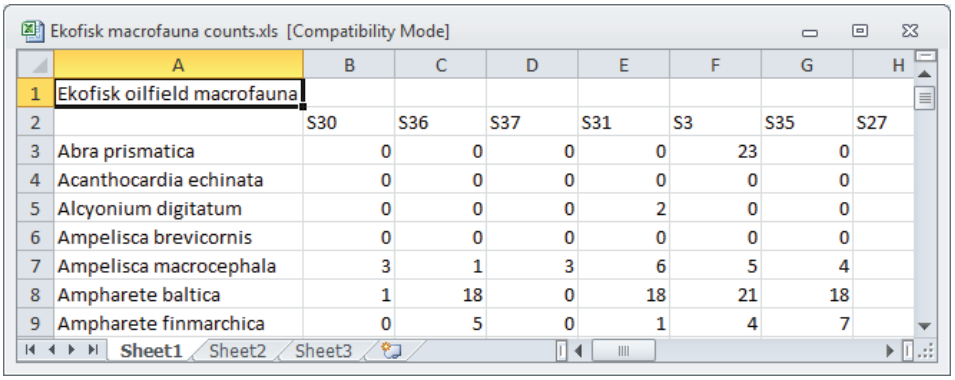
Properties

Edit>Properties produces the Sample Data Properties dialog box, seen earlier, where information about the data can be checked and amended, e.g. Title, Data type, Description, numbers of rows and columns and that the Columns are the Samples in this case. The History box will accumulate information on Pre-treatment operations such as standardisation, transformation, species weighting etc (and for resemblance matrices will specify the coefficient used). **Properties** is also one of the items that can be selected by right-clicking over the data sheet.



Opening Excel files

Usually, rectangular data matrices of variables by samples, or samples by variables, will initially have been entered into Excel. For entry to PRIMER, these should have different data arrays (e.g. abundance, biomass, environmental variables etc) in different sheets – though they can be in the same Excel file – and will need to be read in one sheet at a time. The data format is simple but specific and must be adhered to. For the above Ekofisk counts, an Excel file would have been:


Ekofisk macrofauna counts.xls [Compatibility Mode]

	A	B	C	D	E	F	G	H
1	Ekofisk oilfield macrofauna							
2		S30	S36	S37	S31	S3	S35	S27
3	Abra prismatica	0	0	0	0	23	0	
4	Acanthocardia echinata	0	0	0	0	0	0	
5	Alcyonium digitatum	0	0	0	2	0	0	
6	Ampelisca brevicornis	0	0	0	0	0	0	
7	Ampelisca macrocephala	3	1	3	6	5	4	
8	Ampharete baltica	1	18	0	18	21	18	
9	Ampharete finmarchica	0	5	0	1	4	7	

Sheet1 Sheet2 Sheet3

If referring to the same set of sample locations/times/treatments/replicates, different biotic and abiotic arrays should have the same (unique) sample labels over the different sheets; the samples can then be matched. The label would helpfully be a combination of the particular location/time/treatment/replicate alphanumeric codes, though it could be a simple integer code (1, 2, 3, …). Either way, the place to identify the different *factors* (location, time etc – see Section 2) is not as multiple heading rows at the top of the matrix (only one row of sample labels is allowed) but at the bottom of the array, separated from the data by a blank row. The data array can be entered as the transpose of that shown above (samples as rows rather than columns) but the same principle applies – any factors are placed to the right of the data separated by a blank column (in practice, a blank column *label* is sufficient to make PRIMER believe that the data has finished and the factors started, so you must avoid using a blank sample label). An example of the environmental data for the Ekofisk study, in this transposed format, is now used to step through the input options.

(Ekofisk abiotic data)

For the 39 sites around the Ekofisk oil-field, environmental data is available on concentrations of total hydrocarbons and metals such as Barium, Strontium, Copper in the sediments, and measures of physical sediment properties, such as % mud; also the distance from the oilfield centre is another variable in this array. These are in the Excel file Ekofisk environmental.xls:

Ekofisk environment.xls [Compatibility Mode]

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Ekofisk sediments													
2		Distance	THC	Redox	%Mud	Phi mean	Ba	Sr	Cu	Pb	Ni		Dist#	Dist
3	S30	0.1	232	80	11.91	3.22	1967	587	97	65	14		1	D
4	S36	0.1	443	189	10.95	3.3	1997	509	8	43	10		1	D
5	S37	0.1	1896	85	12.29	3.18	1913	836	22	88	13		1	D
6	S31	0.15	251	124	7.58	3.11	1922	498	14	38	16		1	D
7	S3	0.25	17	142	4.94	3.07	3608	278	5	25	6		1	D
8	S35	0.25	56	153	8.3	3.05	3143	302	5	23	10		1	D
9	S27	0.33	32	150	3.48	2.96	3096	179	4	28	5		2	C
10	S25	0.45	57	168	6.2	3.1	4913	208	4	29	8		2	C
11	S26	0.5	39	213	4.34	2.97	2876	299	5	10	8		2	C

Environmental | Sheet2 | Sheet3

Note that there is only one sheet defined in this case ('Environmental'), it is in the form samples × variables with the same (unique) labels for the sites as in the assemblage data, which can be of any length and consist of a mix of numbers, letters and spaces (as can the title be, if present), but data entries must always be strictly numeric (e.g. '<' signs are not permitted).

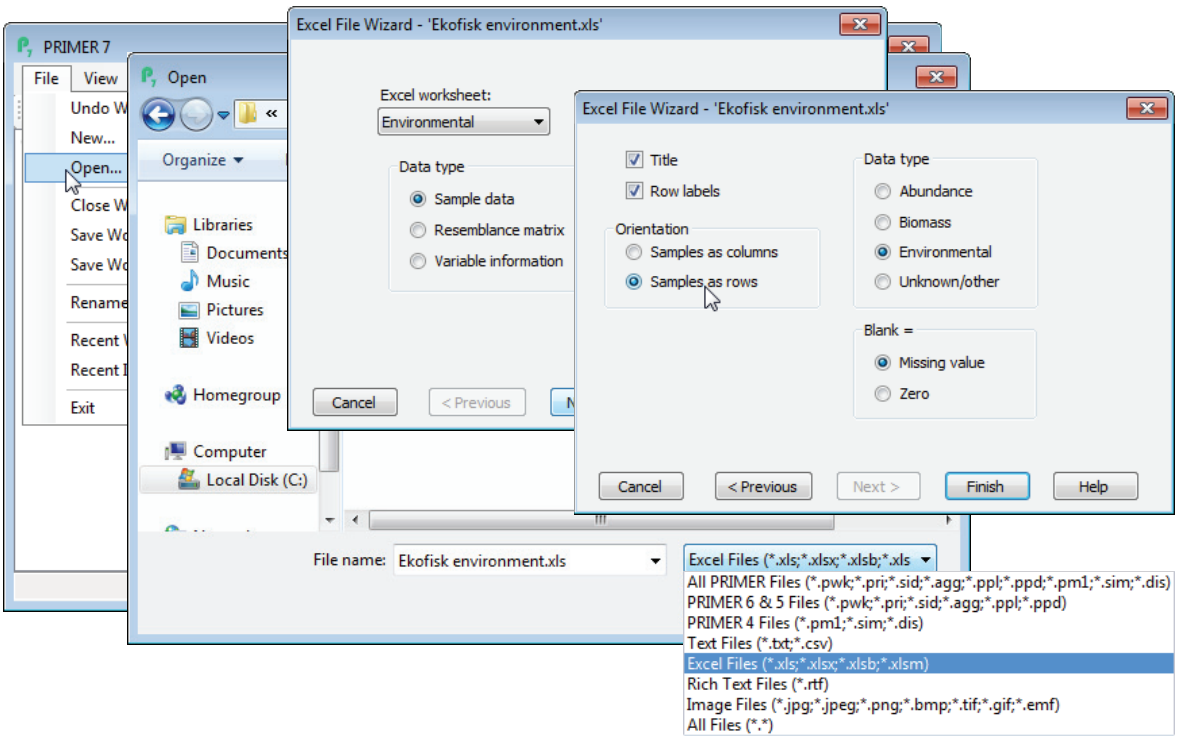
Wizard for input data

File>Open gives the same Open dialog box as previously but, importantly, the drop down menu on the bottom right of this dialog should now be used to select files of type: Excel Files (*.xls, *.xlsx, ..), otherwise the Excel file will not be visible in the file listing. Click on a filename and the **Open** button generates a 'wizard' which guides the user through the choices that must be specified. On the first dialog box, specify which Excel worksheet to use from the file, selected by name, and then what Data type: •Sample data (a rectangular array of samples × variables or variables × samples); •Resemblance matrix (a triangular matrix of pairwise similarities/dissimilarities/distances/ranks/correlations and even pairwise ANOSIM R statistics – PRIMER will itself generate a wide range of these from sample data sheets but it might, for example, be required to input a specialised measure computed by other software into the PRIMER routines); •Variable information (this is a slightly broader category in PRIMER 7 which is still, however, mainly used to hold taxon information on each species, referred to previously as *aggregation files*, in which species are linked to their taxonomic hierarchy of genus, family, order, etc).

v7

On the second dialog box, make sure the Title box is checked (the default option) if there is a separate title line in the top left cell (A1) of the Excel file. Failing to uncheck this box when there is no additional title line is likely to be the commonest source of problems when reading an Excel sheet into PRIMER, the first row of the data matrix then being mistaken for the column labels (also, failing to check the box when there is a title will result in blank input sheet). Similarly there is a check box for the presence of row labels, e.g. species names (which is the default since they are almost always present). Other choices in this dialog specify whether the Samples are to be taken as columns or rows, and whether the data is of Abundance, Biomass, Environmental or Unknown/Other type. This Data type designation is not crucial, but it does allow PRIMER 7 to select natural defaults for analysis choices, e.g. of resemblance coefficient, dependent on specified data type.

To input the above Ekofisk environmental matrix from the C:\Examples v7\Ekofisk macrofauna directory, take **File>Open>**(File name: Ekofisk environmental.xls)>**Open>**(Excel worksheet: Environmental) & (Data type•Sample data)>**Next>**(✓Title) & (✓Row labels) & (Orientation •Samples as rows) & (Data type•Environmental) & (Blank=•Missing value)>**Finish**, as shown:



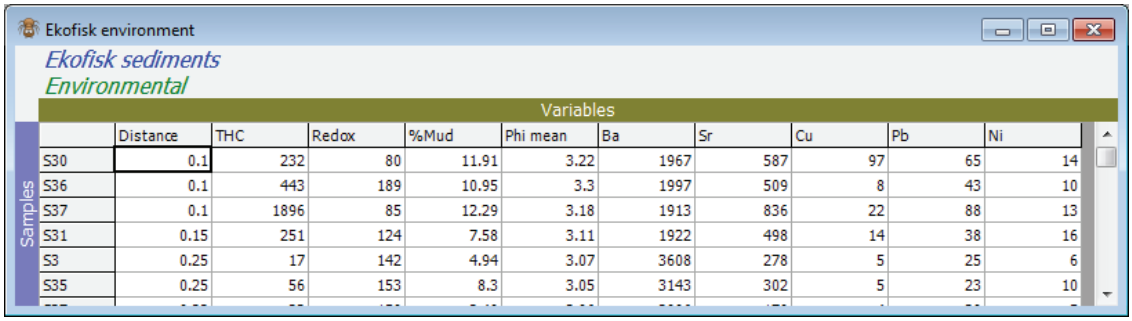
Missing or zero values?

The final option is whether a blank cell in the Excel sheet should be interpreted as a Missing value or a Zero. Typically, it will be Zero for species variables and Missing for environmental or other data. The distinction is important for subsequent analysis: most species-by-samples matrices have large numbers of species that are not present in many samples – they are indicated by zeros, and this information is properly catered for by an appropriate choice of similarity coefficient. If an environmental variable is not detected at a sample site then that should also be recorded as a zero, or as the lower detection limit (or perhaps half that limit). If a specific variable is not measured at a site, through random loss of a sample, then that is properly a Missing value. Inputting a blank cell from Excel, with the (Blank=•Missing value) option, or editing it to a blank after it has been read into PRIMER, will display a Missing! entry.

There are then three possible approaches. For environmental type data which might be transformable to approximate multivariate normality, and for which there are relatively few missing cells, a good option may be to attempt statistical estimation of the (randomly) missing values using the **Tools>Missing** routine. This uses the EM routine to give maximum likelihood estimates of the missing cells by exploiting the correlations among variables (see Section 12), thus completing the matrix. However, in many cases these normality assumptions are not viable, or there are simply too many parameters to estimate. Thus, secondly (and new to v7), PRIMER now automatically takes the simpler approach of calculating resemblance measures after removing, separately for each pair of samples, all variables which have a missing value for either sample. All resemblance measures are then automatically adjusted for the crude bias which results from such *pairwise eliminated data* input to totalled measures, such as Euclidean and Manhattan distance (without this adjustment some pairs of samples would be given greater distance simply because they are summed over more variables), see Section 5. Of course, a third possibility is simply to select a subset of samples and variables for which there are no missing values, e.g. by **Select>Variables>**(•No missing values).

v7

It is important to appreciate that random loss of a whole sample (for all variables), e.g. loss of a replicate community sample from a balanced sampling design, is not thought of as producing missing values. If all species (or variables) are lost for that sample, it is simply omitted, and the design becomes a slightly unbalanced one, which is perfectly well catered for in most of the PRIMER (or PERMANOVA+) routines, e.g. in the ANOSIM or PERMANOVA hypothesis tests.



	Distance	THC	Redox	%Mud	Phi mean	Ba	Sr	Cu	Pb	Ni
S30	0.1	232	80	11.91	3.22	1967	587	97	65	14
S36	0.1	443	189	10.95	3.3	1997	509	8	43	10
S37	0.1	1896	85	12.29	3.18	1913	836	22	88	13
S31	0.15	251	124	7.58	3.11	1922	498	14	38	16
S3	0.25	17	142	4.94	3.07	3608	278	5	25	6
S35	0.25	56	153	8.3	3.05	3143	302	5	23	10

Save the workspace in the C:\Examples v7\Ekofisk directory with **File>Save Workspace As>**(File name: Ekofisk ws.pwk), for later use, and **File>Close Workspace** to clear the workspace. Further files will now be opened from C:\Examples v7\Tasmania meiofauna, to demonstrate text file input.

(Tasmanian meiofauna)

This study concerns meiofaunal abundances from a two-way layout of samples on a sand-flat in Eaglehawk Bay, Tasmania, see Chapters 6, 7 and 12 of the CiMC manual. Separate data arrays are available of nematode and copepod communities associated with disturbed and undisturbed patches of sediment at four locations across the sandflat, the disturbance being caused by natural burrowing activity of soldier crabs (original paper: Warwick RM, Clarke KR, Gee JM 1990, *J Exp Mar Biol Ecol* 135: 19-33). The two disturbance conditions (D and U) are referred to as the *treatments* (though this is an observational study not a manipulative experiment) and the four locations as *blocks* (B1 to B4). For each treatment/block combination there are two replicates. Each replicate is a sediment core for which both nematodes (39 taxa) and copepods (17 taxa) are counted.

Opening several files at once

The directory C:\Examples v7\Tasmania meiofauna contains the PRIMER 7 files of separate nematode and copepod data, Tasmania nematodes.pri and Tasmania copepods.pri. Both can be opened in one step by taking **File>Open** and clicking on these file names, with the Shift or Ctrl key held down. (This operates in the usual Windows way, with Shift-click highlighting all files between the two items selected, and Ctrl-click highlighting individual, non-consecutive files in the list.) Both (in general, all) selected names appear in the File name box and are opened with a single press of the **Open** button.

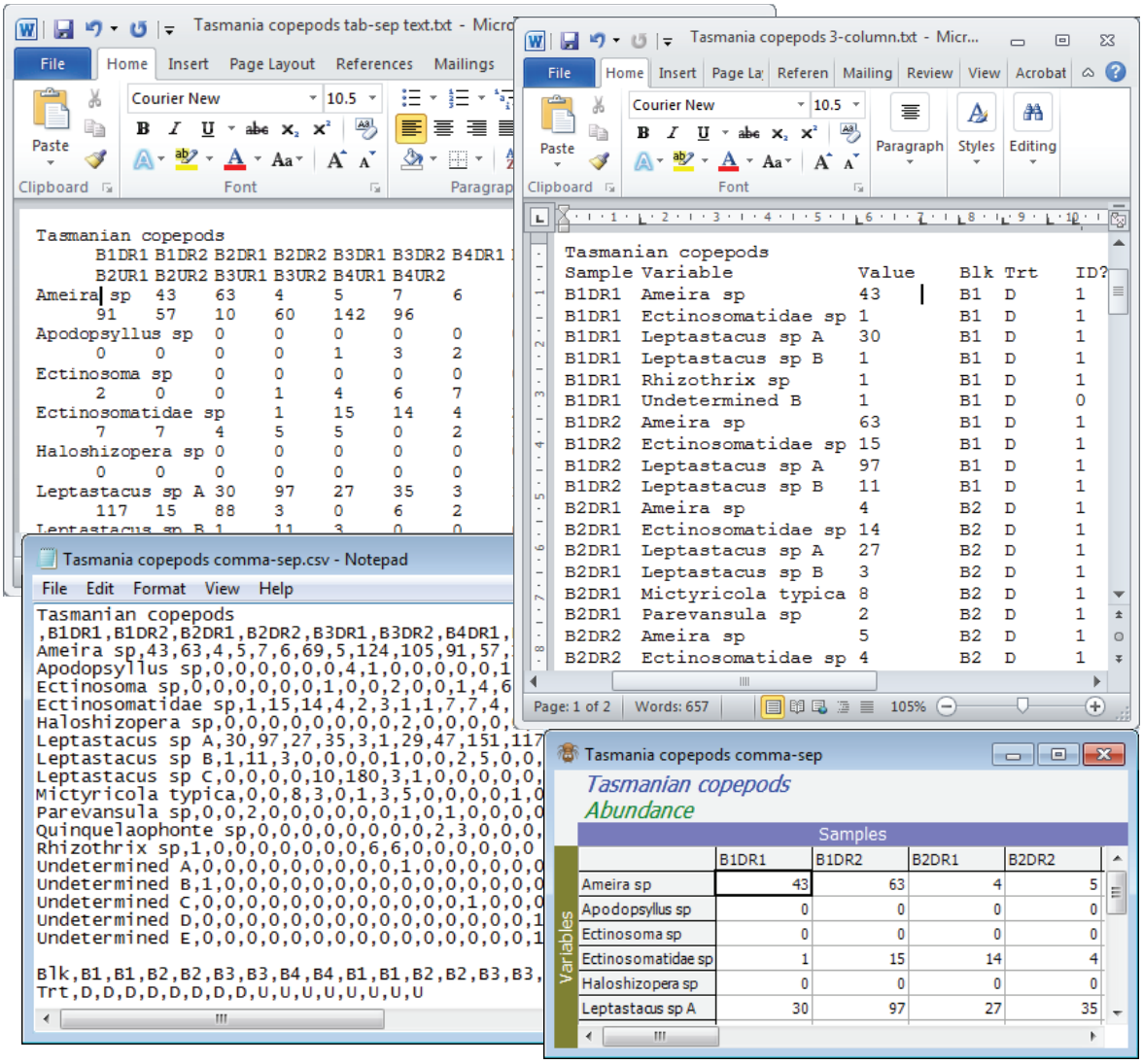
Opening the same file twice

PRIMER 7 will allow the same file to be opened into the workspace more than once, since the response to the **Open** menu item is to create a copy of the file for entry to PRIMER at that time, and there is no physical link maintained from the workspace to the original file. Thus amendment of that external file cannot alter data entries in the workspace, and vice-versa (unless, of course, **Save Data As** is used to overwrite that file with one of the same name), so a second copy of the original file can be opened without difficulty. PRIMER 7 does, however, demand unique naming of all workspace items, so any second (and subsequent) attempts to open the same file will add a version number, e.g. Tasmania copepods(2). Similarly, default names for new windows generated automatically by an analysis sequence: Data1, Data2, …, Resem1, Resem2, …, Graph1, Graph2, … are unique, and if changed to more identifiable names, these too should be unique. Thus an attempt to give the name Tasmania nematodes to a similarity matrix as well as the data sheet will result in the name Tasmania nematodes(2) being assigned.

Text-format input files

The Tasmania meiofauna directory also contains three different text format versions of the copepod samples, Tasmania copepods tab-sep.txt, Tasmania copepods comma-sep.csv and Tasmania copepods 3-column.txt, in addition to a fourth text version of the same data, Tasmania copepods v4.pm1, which is in the original DOS PRIMER v4 format. The first two are rectangular variables × samples (but could be samples × variables) arrays, and differing really only in what is used as a separator (delimiter) between the data entries: *.csv files are comma-separated and *.txt are typically tab-separated (e.g. outputting to *.txt format gives tab delimiters). However, input from *.txt format is more general: it can also cater for comma-separated or space-separated entries or the use of any other specified delimiter. In all cases, rows are separated by (hard) carriage returns but for columns there is no limit on the length of each line, and these will typically be wrapped (with soft carriage returns) when displayed with a text editor or word processor, as seen below. The third file, Tasmania copepods 3-column.txt, is an example of 3-column format, in which each line of the

text file has only three columns of data separated by tabs (other delimiters, such as commas, are also allowed). The format must again be followed exactly: as the second line of header information shows, Column 1 is the sample label, column 2 the variable label, and column 3 the numeric data entry. The advantage of this format is that only non-zero entries need be listed – when PRIMER converts this to rectangular format the blank cells will be automatically filled with zeros (and again without fixed size limits). Importantly, this 'flat-form' structure is the *record format* which many relational databases use to hold observed occurrences or counts of a specific species at a specified location (set the third column to 1 throughout, if these are records only of presence), and the same record format is often also used for abiotic or other measurement variables. All such databases (e.g. Access) will be able to output comma/tab separated text format files of the type shown below right.



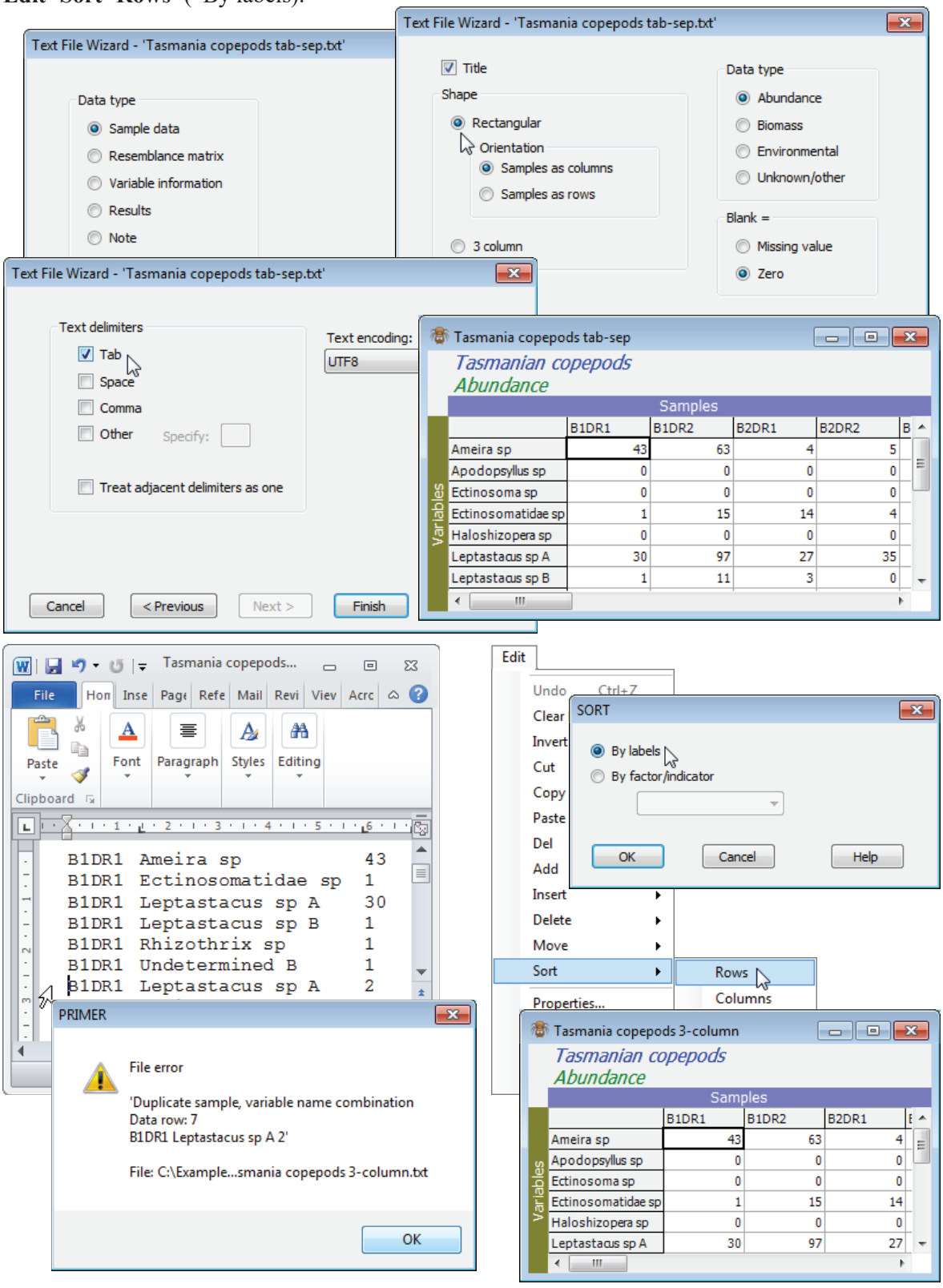
Factors in 3-column text format files

v7

Associated with each record are often one or more factors which define the conditions under which that sample was taken (sites, times etc). These could be copied and pasted from a sample table held in the relational database to a *factors sheet* set up for that data in PRIMER – see section 2 for how to set up factors within PRIMER – but this categorical information on the data structure is typically output from the relational database as part of each record. If there is no access to the original database table of factors, such a record format can be used by PRIMER 7 to populate its factor table. This is why the above 3-column example does not contain just three columns, in this case! There is a fourth blank column (i.e. an extra tab) and then two (alphanumeric) factor level columns – though there could be many more – which here define the block and disturbance status for each record. Of course each combination of the two factor levels is repeated as many times as there are numbers of species observed in that sample, and if these entries are not identical an error message will result. Then (if needed) follows another blank column (extra tab) and any 'factor levels' defined on the variables (termed *indicators*, see Section 2); here this indicates which species have been identified.

Dialog for input of text format files

Read in the first of the above text files: **File>Open>**(Files of type: Text Files (*.txt, *.csv)) & (File name: Tasmania copepods tab-sep.txt), using the Text File Wizard dialogs shown below. Repeat for Tasmania copepods comma-sep.csv, the only difference being the switch to (Text delimiters ✓Comma) in the third dialog box. For Tasmania copepods 3-column.txt, the below shows the error message obtained having inserted at line 7 a mistaken repeat of the same combination, '`B1D1R1 Leptastacus sp A 2`'. When the error is corrected, the Wizard dialog proceeds exactly as for the other two cases, except that the option (Shape•3 column) is selected in the second box. Note that the 3-column file enters PRIMER with a different species order (the order in which variable names are encountered in the file); it will coincide if rows are re-ordered alphabetically with **Edit>Sort>Rows>**(•By labels).



Size of data worksheets

There are no fixed size limits for arrays within PRIMER 7, simply an overall limit determined by the amount of available real memory on the computer. There will, of course, be significant time constraints for some of the more compute-intensive routines, and there is a limit to how many samples it is sensible to try and view at the same time, in ordination plots for example, but it is a viable strategy to place all related data into a single workspace, and data of the same type (e.g. counts for a specific faunal assemblage in a complete set of samples) into a single worksheet. Defining a structure of factors on the samples (see Section 2) will allow the selection of subsets from that worksheet, or averaging over replicates (or factor levels), needed for a specific analysis. Continually improving computational power makes it possible at least to hold and manipulate arrays with thousands of species (typically OTU's in microbial applications) and thousands of samples, even if successful analyses will often involve targeted selections or averages of samples. Whether matrices are input as samples (rows) × variables (columns) or variables × samples is not of relevance: PRIMER simply needs to be told whether the samples are the rows or the columns. Note that transposing a worksheet within PRIMER is possible (by **Tools>Transpose**), but will not correct a mistaken attribution on entry – if the rows have been incorrectly called 'samples' during the **Open** dialog then, after transposing the worksheet, the columns will still incorrectly be called 'samples'. Instead, the mistake is corrected by **Edit>Properties** and taking Samples as•Columns.

Merging worksheets

For data collection reasons, it may still be the case that data from essentially the same array are sourced from several different file, e.g. abundance of comparable species lists over a set of sites but with data from different years held in different Excel sheets. If those sites, or some of them, need to be analysed over the years then the best strategy is often to read all the different files into PRIMER and then **Tools>Merge** them to a common worksheet. Before entry, the sample labels should be unique, e.g. identify the year as well as the site and the replicate number etc, and species names (or numbers) be consistently spelt. Then runs of **Merge** will stitch the sheets together, a pair at a time, expanding the species lists accordingly to take account of the fact that different years may have species lists of slightly different composition, length or order, and zeros will be added in relevant cells (or Missing! if this is selected as more appropriate, e.g. as it would be for environmental data).

Not all data for the same set of sample labels should necessarily go into a single worksheet, e.g. species abundances and environmental variables for the same set of sites/times are usually best kept in separate arrays because sample resemblance matrices (Section 5) will often need to use different coefficients. Whether environmental information is best held as a separate data array or as a factor sheet associated with a species array depends on the data type and context: factors are categorical (whether unordered or numerically ordered) whereas data arrays are numerical. Some variables may appear in both ways, e.g. water depth in an abiotic matrix and as a factor (shallow, mid, deep).

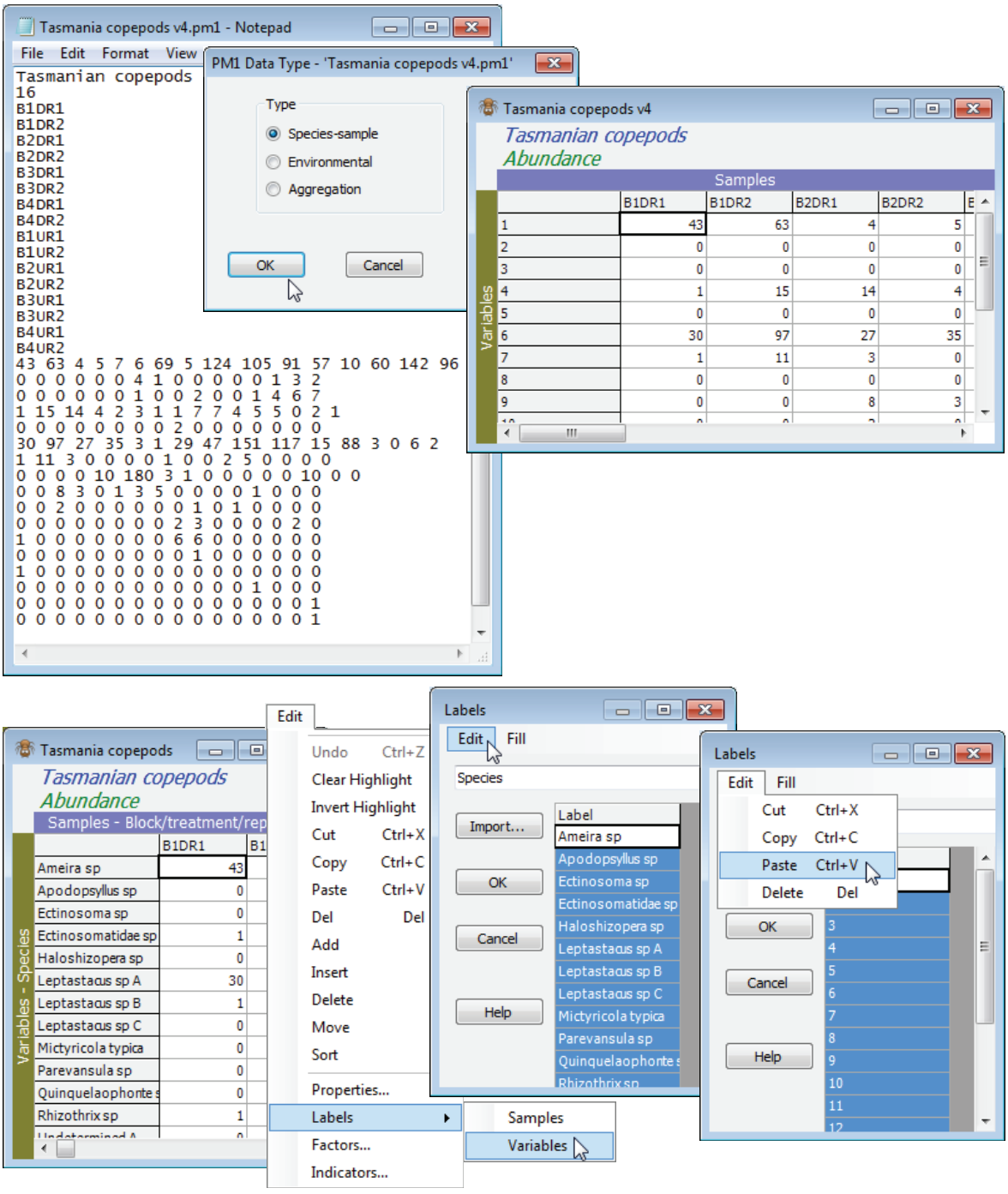
Output data formats

Output format options, with **File>Save Data As**, are generally the reverse of input choices. The default is a PRIMER 7 (binary) file but data sheets (or resemblance matrices) can also be saved in earlier PRIMER 5 and 6 binary formats, and to Excel in current *.xlsx format or the older *.xls (with its restriction to < 255 columns). Text files can also be output in either rectangular or 3-column format, the separators then always being Tabs. The very early DOS PRIMER 4 text format can also be output (or input) but much of the associated information (e.g. species names, factors) is lost, and this format is likely only to be of interest at this stage in restoration of old archives. An example of the PRIMER 4 format can be seen in the file Tasmania copepods v4.pm1 below, which contains the same counts as Tasmania copepods.pri though cannot hold species names or factors in the same file. (Note that line 2 defines the number of samples, followed by the sample labels and then species counts separated by any combination of single or multiple spaces or tabs).

PRIMER 5 files did not have a defined data type (so are all read in by PRIMER 7 as Data type• Unknown/other). They also had no History box (defining standardisations, transformations etc which may have been applied to obtain the current sheet). It follows that outputting PRIMER 7 files in PRIMER 5 format will lose the information on Data type and History. As noted at the start of the section, in contrast, differences between PRIMER 6 and 7 data formats are rather minimal (though workspace files are very different). However, PRIMER 7 is not backwards-compatible in general, so that PRIMER 6 cannot open files created in PRIMER 7 unless they have been explicitly saved to the earlier format. Naturally, PRIMER 7 is forwards-compatible and will automatically open any earlier data or v6 workspace files.

Editing labels

Take **File>Open**>(Filename: Tasmania copepods v4.pm1)>Type•Species-sample to input this (archival) v4 format file, and note that the missing species labels could be copied and pasted from elsewhere (if they were available in the same order) – perhaps an external file or, as demonstrated below, from another worksheet within the workspace (Tasmania copepods.pri). All amendments to labels are implemented through **Edit>Labels>Variables** (or **Samples**), then clicking on the label header will highlight the full set of labels for copying out – or pasting into – their contents.



Save the current workspace in the C:\Examples v7\Tasmania meiofauna directory, with **File>Save Workspace As>**(File name: Tasmania ws.pwk), for use in the next and later sections.

2. *Factors* (and *Indicators*), identifying sample (and species) groups

Active window

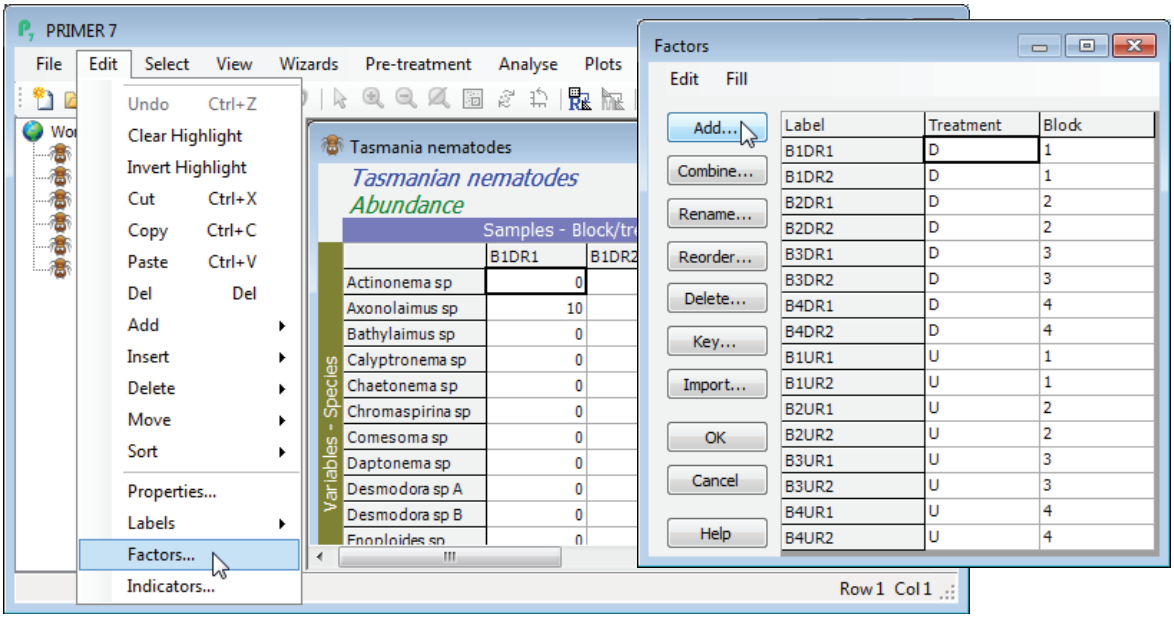
If you have been carrying out the manipulations in Section 1, by now you will have several sheets open in the C:\Examples v7\Tasmania workspace, the worksheet Tasmania nematodes and several identical versions of the copepod assemblages. Unclutter your PRIMER desktop by **Window> Close All Windows** and then re-display just Tasmania nematodes and Tasmania copepods by clicking on their icons in the Explorer tree. (If the workspace is clear, then **File>Open** these two *.pri files). It is fundamental to operation of PRIMER that only one window in the workspace is considered *active* at any one moment, and this will always be displayed on the PRIMER desktop and be identified by the slightly deeper colour title bar and the highlighted entry in the Explorer tree. You can select which one to activate by clicking anywhere on its window or the entry in the tree. Menu selections apply only to the active window, e.g. **File> Save As**, **Edit>Labels**, and the **Analyse** and **Tools** items (though these may specify one or more secondary sheets needed for a composite analysis). Note also that menus are dynamic, with content that changes with the context. When the active window is a rectangular data sheet, different **Analyse** options (e.g. Resemblance, DIVERSE, PCA) are available than for a triangular resemblance matrix (e.g. CLUSTER, MDS).

Use of factors

With Tasmania nematodes as the active window, select **Edit>Factors** from the main menu (or use the shortcut right click when the cursor is over the data matrix to bring up a combination mainly of the Edit and Select menus), and observe that there are already two factors defined. The treatment factor *Treatment* splits the 16 Tasmanian sandflat samples into two levels, namely whether they are from disturbed (D) or undisturbed (U) areas of sediment. The *Block* factor divides the samples up in a different way, into four levels, the four separate sampling patches across the sandflat (1 to 4). In statistical terminology, the treatment and block factors are crossed, meaning that there are samples at every combination of levels of the first and second factors. Factors are heavily used throughout PRIMER, in at least two main ways:

a) to define a group structure for multivariate hypothesis testing (e.g. ANOSIM, see Chapter 6 of the CiMC manual). Such *a priori* structuring of the samples (i.e. prior to seeing the data) plays an important role in formal inference about sample patterns, and also the interpretation of which variables (e.g. species) are primarily responsible for distinguishing specific groups (Chapter 7);
b) purely as a means of labelling points on plots, in dendrograms etc, in which case there might be a different 'level' for every sample, e.g. a fuller or more abbreviated site name than is held in the sample label. There is no limit on the length or alphanumeric content of a factor level.

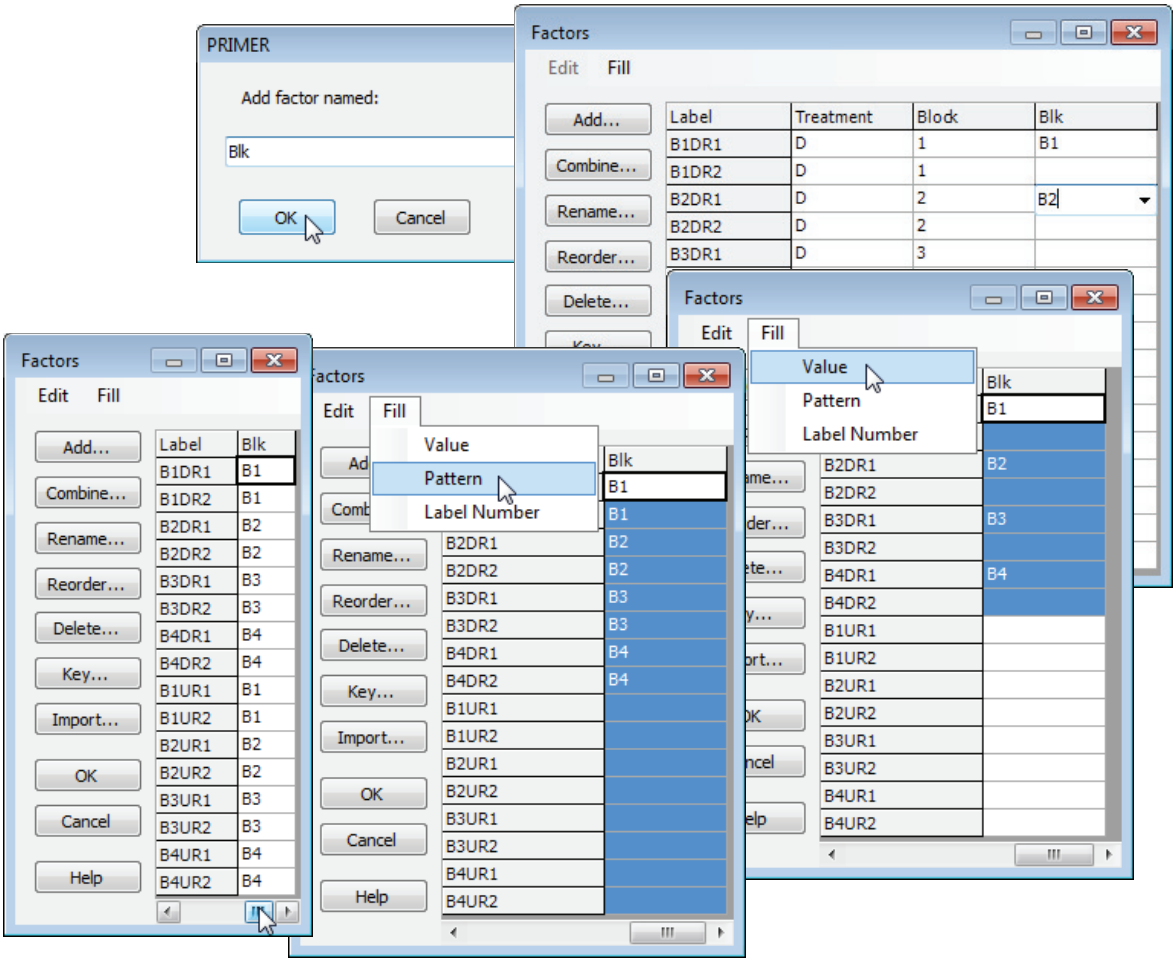
Factors are carried around and saved as part of the data sheet they are linked to, and not saved as separately named data sheets. This is in contrast to (numeric) environmental variables associated with each biological sample, which are held in a separate sheet – preferably with the same sample labels as the biota, and which could have some or all of the same (categorical) factors defined. To emphasise that block designation is purely a category here, not a numeric sequence, a new factor *Blk* will added here, with levels B1, B2, .. not 1, 2, … (as seen in the previous text file versions).



Creating & filling in factors

v7

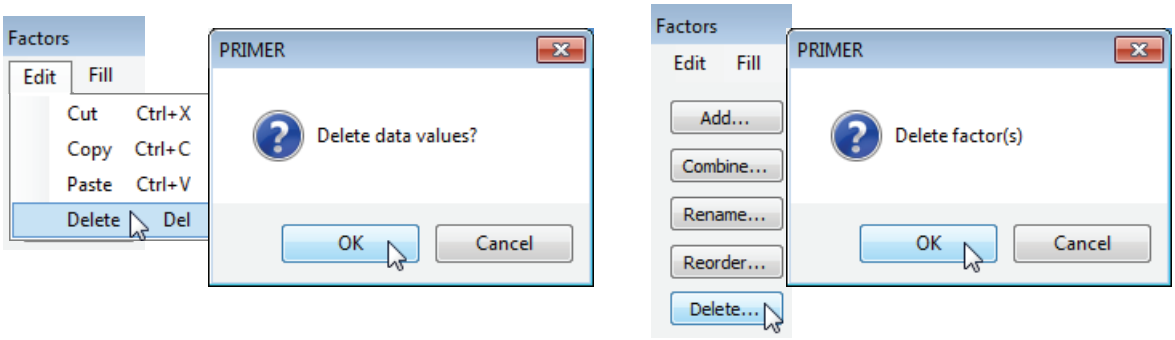
In the Factors dialog box (obtained from **Edit>Factors** on Tasmania nematodes) take **Add>**(Add factor named: Blk). The cursor is then at the top of the new (blank) label column, ready to start typing. You need only put in the first entry for each new level (B1, B2, ..) if they are in groups of identical values. (There are only two replicates per cell here so only pairs of identical values, and it is just as quick to type them all, but this new feature in PRIMER 7 will typically save much typing of identical strings, so is demonstrated below). Having entered B1, B2, B3, B4 in the relevant rows (1, 3, 5, 7), highlight the first 8 entries and take **Fill>Value**, which fills in the blanks in the top half. (For any run of blank entries in the highlighted area, **Fill>Value** will simply repeat the last filled entry immediately above them). The same sequence then needs to be generated for the second set of 8 entries and this is quickly achieved by highlighting the whole column, clicking on its label (*Blk*), then taking **Fill>Pattern**. (This copies any run of fully filled entries into any blank entries starting immediately below them, stopping part way through if necessary, if it gets to another filled entry – filled entries are never overwritten – and then repeating this through the highlighted area).



Cut, Copy, Paste, Delete in factors

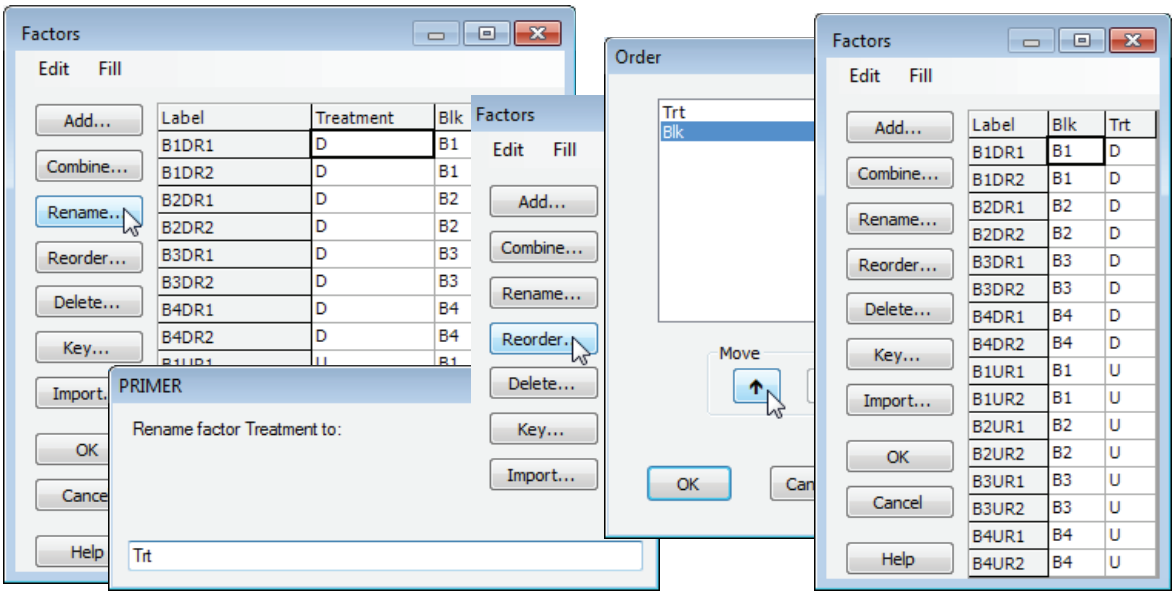
An alternative (and clumsier!) way of creating this factor would be type in the top half, then highlight and **Edit>Copy** these 8 entries and **Edit>Paste** them when the cursor is at the start of the lower set. (Pasting does, of course, overwrite existing entries, as in normal Windows practice). The usual **Cut**, **Copy, Paste** and **Delete** operations can be performed with key strokes (Ctrl-X, Ctrl-C, Ctrl-V and Del key), rather than from the **Edit** menu, in a fully standard way throughout PRIMER. Deletes will trigger a query of 'Delete data values?' because there is no Undo option on the Factors dialog. If a significant deletion takes place accidentally, the best strategy is to abort any changes made to the factors sheet since it was entered on this occasion, by the Cancel button (which again throws a query box of 'Cancel all changes?'), and you can reopen the factors sheet and try again.

Do not confuse deleting factor entries with removing the whole factor (or several factors) from the factor sheet. The latter can be achieved by highlighting the factor – clicking on its label at the top of the column (or clicking and dragging to capture several consecutive factors) – and taking the **Delete** button on the left of the dialog box. Try this out by deleting the (now redundant) *Block* factor. This also generates a query box, this time of 'Delete factors?' not 'Delete data values?'.



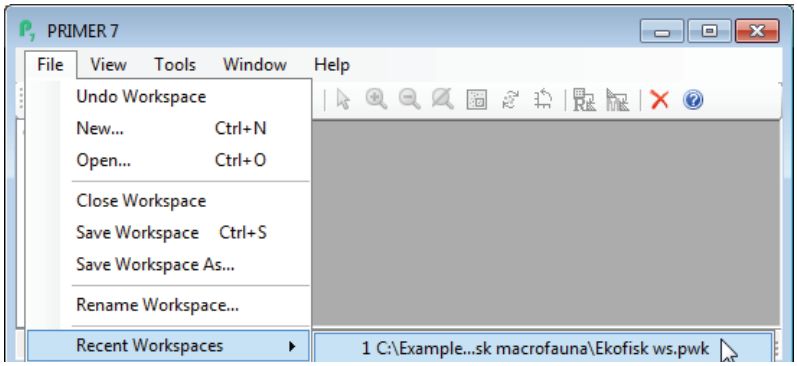
Renaming & reordering factors

Finally, to make factors in the Tasmania nematodes sheet consistent with the text format copepod files of Section 1, rename the Treatment factor as *Trt* using **Rename**>(Rename factor Treatment to: Trt), and rearrange the order of factors to put *Blk* first, with **Reorder** clicking on Blk and Move↑.



Multiple sessions and recent workspaces

As a further example of **Fill>Value** to quickly set up a factor of group levels you might like to re-open the saved workspace from the oil-field study of Section 1, Ekofisk ws.pwk. Taking File>Open and supplying the workspace name from the directory C:\Examples v7\Ekofisk macrofauna would lead to a prompt to save the currently active Tasmania workspace prior to shutting it down, in order to open the Ekofisk workspace. If it is useful to have both sets of data being open and worked on at the same time (though independently), a different solution is needed. This is generally <u>not</u> to open Ekofisk data files into the Tasmania workspace – data sets that will never interact in a common analysis are best kept in separate workspaces – but to launch multiple runs of PRIMER. These will not interfere with each other: a copy is taken of the current version of each file at the time it is loaded into the workspace, so the original file is never then modified by internal workspace actions or saving the workspace (only explicitly taking **Save Data As** and providing the same file name can alter the original file's contents, and even this requires a confirmation stage before it is over-written). In this second PRIMER desktop therefore, re-open the Ekofisk workspace with **File> Recent Workspaces**>C:\Examples v7\Ekofisk macrofauna>Ekofisk ws.pwk.

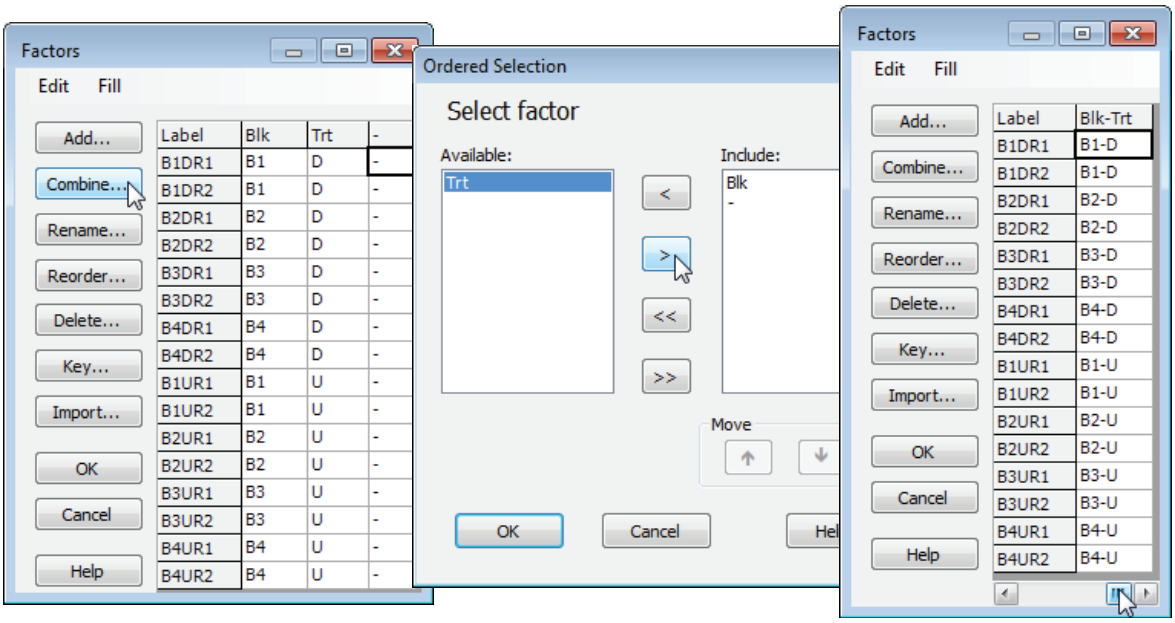


Now take **Edit>Factors>Add**>(Add factor named: Dist#) in order to match the alphabetic codes for the different distance groups of sites from the oil-field centre: D, B, C, A with numeric ones: 1, 2, 3, 4 respectively. (This will be needed for a later example when it is useful to treat this factor as ordered categorical – PRIMER does not treat alphabetic levels in factors as providing an ordering). *v7* As previously, only a 1 in the first row (site S30), a 2 in the S27 row, a 3 opposite S4 and a 4 at S18 need to be typed in, then highlight the column and **Fill>Value**. Another simple example of the use of **Fill** would be to produce a continuous ordering of distances from the oil-field centre, by adding a new blank factor *Dist order*, then highlighting and filling it with **Fill>Value>Label Number**, since here the sample rows in the file have already been ordered by increasing distance from the field.

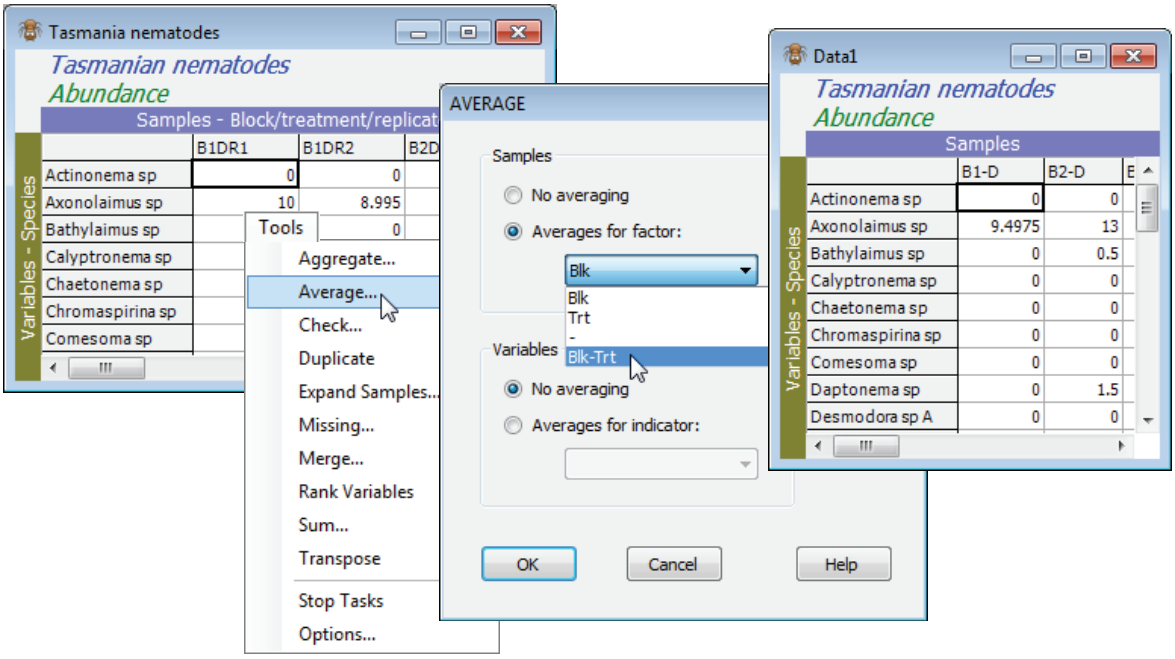
Resave the workspace by **File>Save Workspace** – there is no prompt for a workspace name this time since it has already been saved as Ekofisk ws (as can be seen from the top line of the Explorer tree). Saving to a different workspace name requires **File>Save Workspace As**. You can now exit this PRIMER session with **File>Exit** but note that the previous PRIMER session on the Tasmania meiofauna data remains open, and we will use this to look at the remaining Factor dialog options.

Combining factors (e.g. to average)

With the Tasmania nematodes sheet active, open the Factors dialog with **Edit>Factors**. Combining factors (**Combine**) can be a quick and effective way of creating new factors or composite sample names in nested or crossed layouts. Firstly, though, it is usually useful to create a separator 'factor' (or perhaps more than one), by **Add**>(Add factor named: -), filling the column with dash symbols, by entering a dash in the first row, highlighting the factor and using **Fill>Value** again. **Combine** now displays a typical selection box (PRIMER uses a similar dialog for many other analyses, e.g. selecting a subset of the data by levels of a factor). Click on Blk and [>] then - [>] and Trt [>], to set up which factors are to be combined and in what order. (Note that the double arrows [>>] move all items from the (Available) list to the (Include) list, or back, and a selection of entries can be moved in one operation by holding the Ctrl key down as the items are clicked – or the Shift key to obtain a range of items – as in usual Windows practice). Pressing **OK** then gives a composite factor with name *Blk-Trt* and the 8 levels: B1-D, B2-D, …, B4-D, B1-U, …, B4-U, which are the 8 cells of the two factor crossed design, with two replicates at each level.



Such a combined factor has several uses, e.g. it can be a composite label on an ordination plot, and it is essential for averaging over the replicates in the data, to obtain a matrix of mean values, for each of the 8 block × treatment combinations here. This is simply achieved with an **OK** for all the changes you have made to the Factor information, and then **Tools>Average**>(Samples•Averages for factor: Blk-Trt) & (Variables•No averaging). This creates a new data sheet, Data1, in which the sample labels are the levels of the combined *Blk-Trt* factor, as seen above (B1-D, B2-D, etc). It also carries across what factor information it can from the original sheet (take **Edit>Factors** on Data1), though a factor for which different levels have been averaged over will have 'Undefined!' entries (e.g. produce averages for factor *Trt*, and the *Blk* factor entries would all be undefined, naturally).

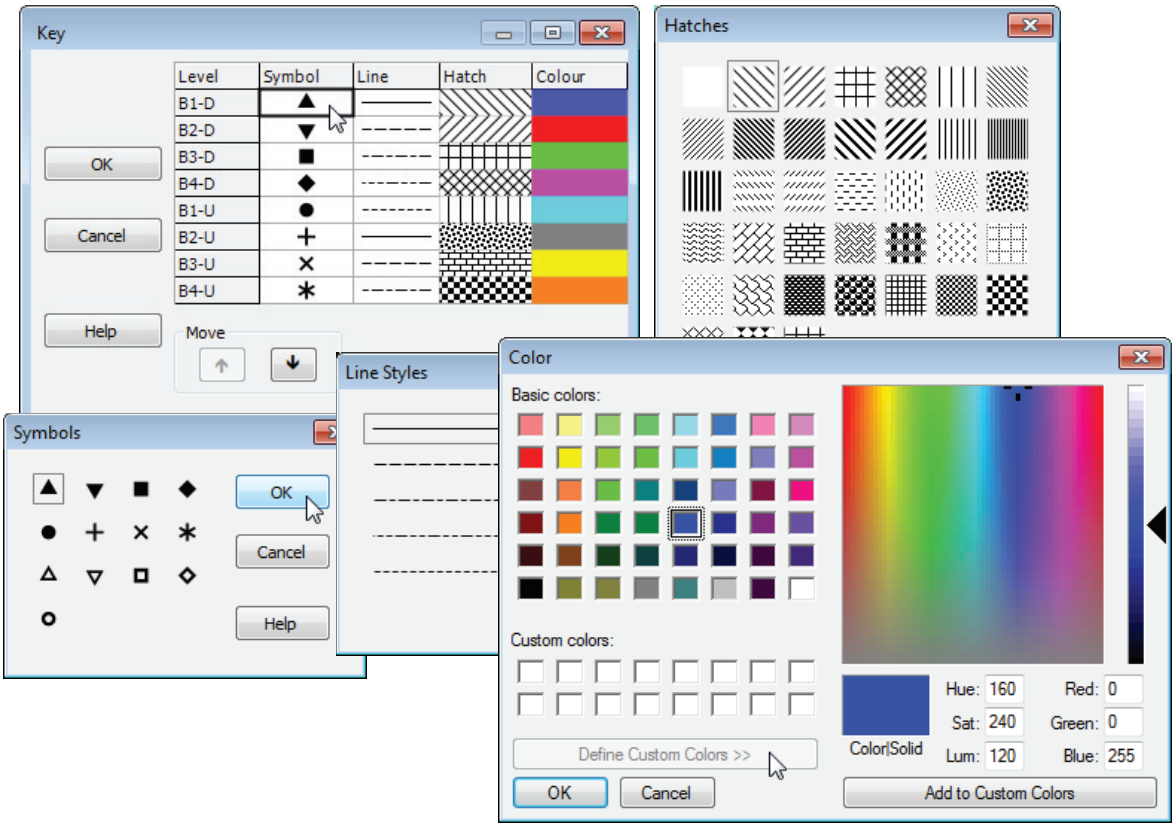


Factor keys

A further button on the **Edit>Factors** dialog box is **Key**, which you could examine with the factors for the Tasmania nematodes. With the cursor somewhere on the combined *Blk-Trt* factor, clicking **Key** gives a display of symbol type and colour for each of the 8 factor levels that will be used on ordination plots etc, and also the line style for joining points (e.g. in dominance curves, see Section 16). Any of these (local) defaults can be changed by double-clicking on one of the cells of the

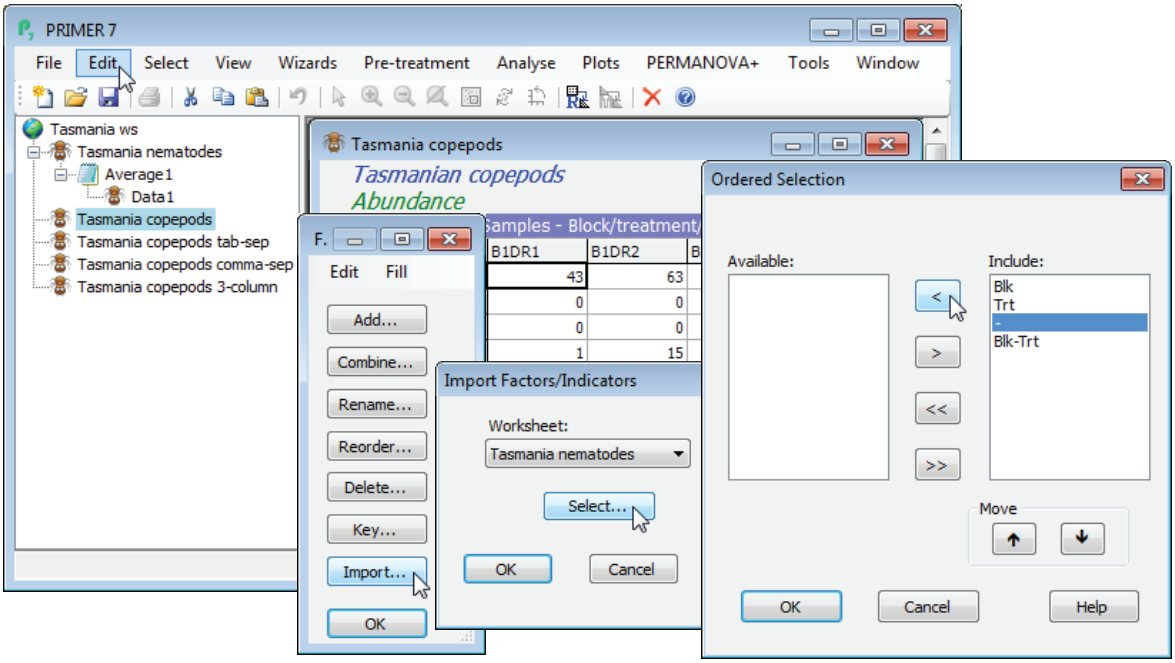
v7

display: on a colour gives a colour chart (48 basic and many custom colours), on a hatching symbol gives 38 fill characters for mono plots, on a symbol gives 13 symbol shapes, and on a line 5 line styles. Key changes will only apply to the specific factor; they can be made in advance, or on the plots themselves. Changes propagate forwards through derived windows, and (usually) backwards to those that are precursors to the current window. The Key dialog is also where factor levels can be placed into the desired order for presentation as a key to symbols on MDS plots, dendrograms etc. To do this, use the Move↑ buttons, up or down repeatedly, on each selected level at a time.



Importing factors

New factors can be created at several stages during an analysis, not just when the active window is a data sheet (e.g. from a resemblance matrix or even a plot) and the new information is propagated both forwards and backwards through the same branch on the Explorer tree. (There are exceptions to backward propagation, in cases where an action, such as **Tools>Average** or **Sum**, fundamentally restructures the samples – existing factors are propagated forward through these steps but not back, understandably). However, when two sheets are in the same workspace but otherwise unconnected (e.g. they are on branches from different initial data sheets), factor information which applies to the sample label names which they share can be transferred between them using the **Import** button on the Factors dialog. An example is the Tasmania copepods(.pri) sheet, which should already be open in the Tasmania workspace. **Edit>Factors** shows that it currently has no factors defined, but its samples (and, importantly, their labels) are identical to those for the nematode data sheet. Taking **Import>**(Worksheet: Tasmania nematode) & (**Select**) gives a selection box, which should list the three factors that were created for the nematode data: *Blk*, *Trt* and *Blk-Trt*. Any factors that are not needing transfer are excluded just by moving them, with <, from the Include: to the Available: box (you might like to do this with the separator column '-'). Then take **OK** on this and the next two boxes, and the desired transfer of three factors to the Tasmania copepod data sheet will occur.



Label matching

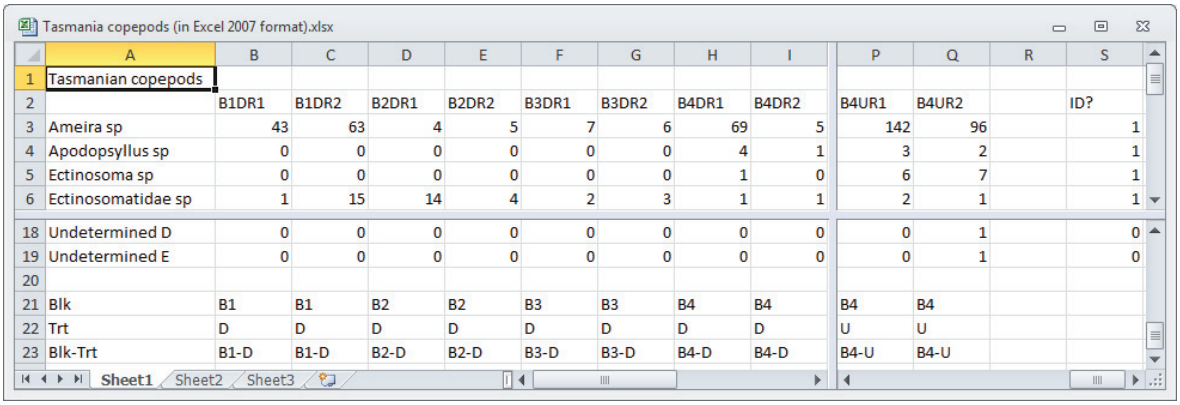
Alternatively, the same endpoint could have been achieved by Adding three new blank factors to the copepod sheet and copying and pasting the contents of the *Blk*, *Trt* and *Blk-Trt* columns from the nematode factor sheet. If importing entries from an external source, such as an Excel column, this approach may sometimes be necessary but it is only appropriate when the samples are in the same order in the two data sets (as they are here). In contrast, **Import** operates by matching up the sample labels in the two files and can therefore re-order the factor levels appropriately when the samples are in a different order. This is a general feature of PRIMER 7 – a lot of use is made of label matching across data sets in this way, which is why it is vital that labels are defined uniquely within a set and carefully checked for consistency of spelling across sets. Of course, if the two sets of sample labels are not identically defined, but do refer to the same set of samples, in the same order, then a copy and paste of the factor content is the only way of transferring the factors.

Factors in *.xls(x) or *.txt files

As noted in Section 1, factors can be created as part of the Excel or text files which are the usual means of inputting data to PRIMER 7. Similarly, data sheets that are saved from PRIMER to Excel (*.xls or *.xlsx) or text (*.txt) formats will automatically export the factors also. The principle is that, when the data has samples as columns, any factors are placed in the input or output sheet as additional rows at the bottom of the array, separated from the data by a blank row. If samples are rows, factors are held as columns to the right of the array, again after a blank column. The 'record' text format differs slightly: after the 3-columns (sample label, variable label, data value) comes a blank column and factor levels (then possibly a blank column and *indicator* levels – see below).

v7

Take **File>Save Data As**>(Save as type: Excel 2007 Files (*.xlsx)) to output Tasmania copepods in Excel format, and open Excel to examine the form in which factors are output (and input). Text format versions of the same data (with *Blk* and *Trt* factors only) are shown in Section 1.



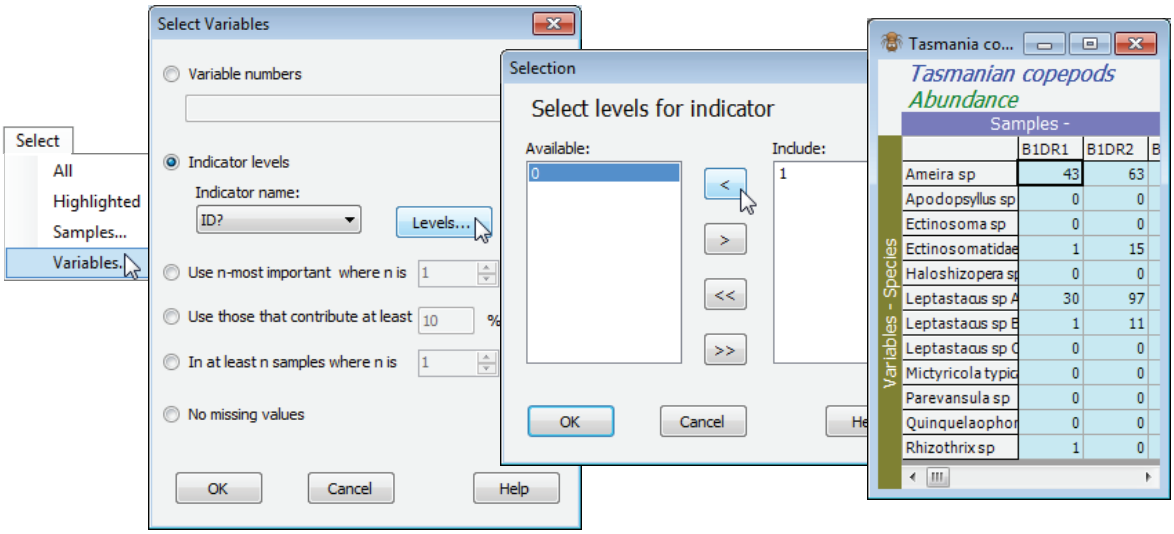
Creating indicators on variables

Indicator is the term PRIMER uses for a factor defined on the variables not on the samples. It is convenient to use a separate term because 'factor' has a well-established statistical meaning (e.g. in ANOVA-type layouts), and refers to structures defined on samples, not on variables. Indicators are less used than factors in practice, though they have a useful role in selecting or removing subsets of variables for the analysis of samples (e.g. only analyse the metals data rather than all environmental variables; only analyse zooplankton, omitting phytoplankton species etc). Adding and manipulating indicators, however, proceeds exactly as for factors, with parallel choices of Add, Combine, Rename, Reorder, Delete, Key and Import from the Indicators sheet produced by **Edit>Indicators**.

A simple example is seen in the Tasmania copepods(.pri) data sheet. **Edit>Indicators** (also on the right click menu when the cursor is over the data sheet) shows the indicator *ID?*, which records whether each taxon has been identified (1) or is an undetermined specimen (0). The *ID?* indicator is also shown above (far right) in the Excel format of this file.

Indicators in selection

Selection by indicator levels is demonstrated by **Select>Variables>**(•Indicator levels)>(Indicator name: ID?)>**Levels>**(Include: 1) & (Available: 0), giving a subset of the Tasmania copepods data sheet which drops the undetermined species. Of course, for such a small data set there are simpler ways of dropping these last five species – see the range of selection options in Section 3.



Now reverse the selection by **Select>All** (and **Edit>Clear Highlight** if you wish), and resave the Tasmania ws.pwk workspace, using **File>Save Workspace**, for use in later sections.

One apparently obvious application for indicators is to specify which species belong to which higher-order taxonomic groups. If separate multivariate analyses are required by major phyletic group, for example, then the different phyla should be set up as an indicator on the species, since this will allow easy selection of the species in a single phylum from the samples × species sheet.

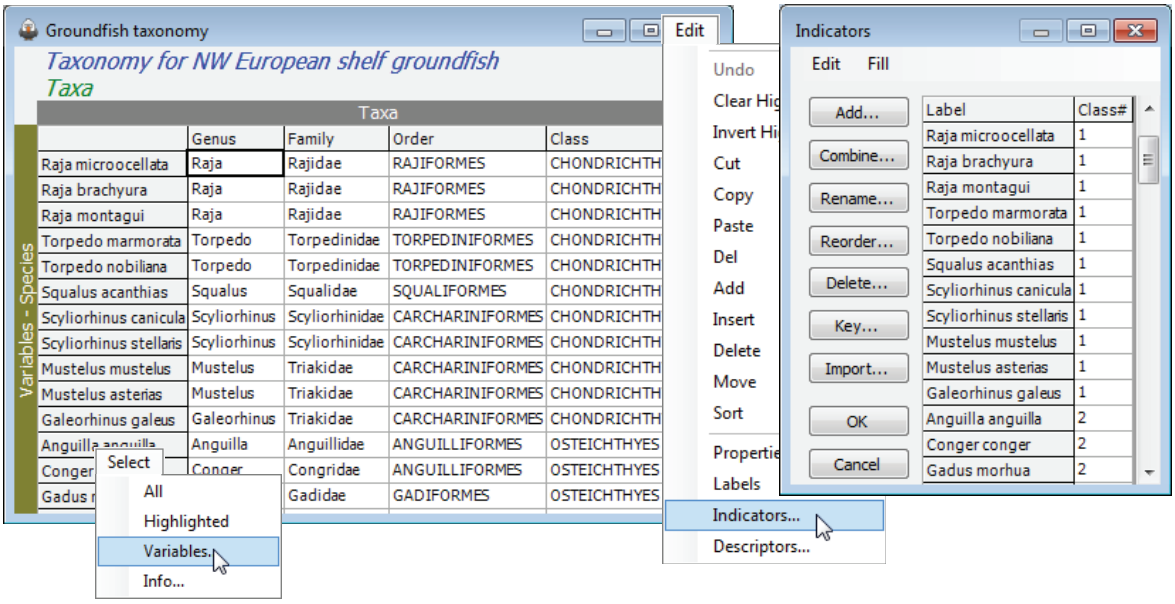
Variable information (aggregation files)

However, the full range of hierarchical indicators represented by a Linnaean classification (which species belong to which genera, which genera to families, families to orders, etc) are usually also best held separately, as a different type of array – that of *variable information*. Mainly for historical reasons these are termed 'aggregation files' in PRIMER, since their initial use was for aggregating species abundances up to genus, family, order, … level information, to judge the extent of change to the interpretation of analyses under coarser identification of taxa (see Chapter 10 of the CiMC manual), and this binary file format is therefore denoted by *.agg. However, in PRIMER 7, arrays of variable information can be more general (and have other Type definitions than *Taxa*). Former aggregation file formats can be opened and PRIMER 7 outputs the full range of previous types, e.g. Save as type: PRIMER Var Info Files (*.agg) for PRIMER 7 (binary); PRIMER 6 or 5 aggregation files.agg (also binary); and simple Text (*.txt) or Excel (*.xls) (or *.xlsx) sheets. Examples using aggregation files will be seen later (Sections 5, 11, 15) though the simple rectangular format is seen here by opening Groundfish taxonomy.agg from the C:\Examples v7\Europe groundfish directory. Three ways in which it might be used are to: a) aggregate abundance to higher taxa with **Tools> Aggregate** (Section 11, and Chapter 10 of CiMC); b) compute biodiversity indices based on the relatedness of species in a single sample, e.g. with **Analyse>DIVERSE** (Section 15, Chapter 17); c) compute resemblance measures between two samples reflecting (higher) taxonomic relatedness of the species found there (Section 5, Chapter 17).

v7

v7

This new *variable information* sheet (below) permits the non-numeric entries which are essential for a variables × taxa 'look-up' table but also, and newly in PRIMER 7, will carry over several of the general features of sample × variables arrays, in that indicators defined on the variables can now be carried around with this array. This might permit the aggregation file to hold alternative names for single species, for example, with an indicator that can be used to select only the taxonomic revision relevant to the historic date of collection/identification of the species count matrix. Importantly, it also allows easy selection of aggregation file subsets, e.g. for testing taxonomic distinctness indices against differing 'master lists' by region, habitat or faunal group (Chapter 17 of CiMC). The simple indicator in the example below could be used to select only the Osteichthyes (Class# = 2) from the Variable information: Groundfish taxonomy, as well as from the data: Groundfish density(.pri).



Note the final entry on the Edit menu here. The concept of *Descriptors* is not particularly relevant to Variable information of type Taxa (they are potentially more relevant to other types of Variable information) but they are the third construction logically needed. Categories (or alternative labels) applied to Samples are termed *Factors*, when applied to Variables they are called *Indicators* and when applied to Variable information they are *Descriptors*.

3. Highlighting and selection (*Select*)

Highlight *and* select

There are many cases in which analyses of different subsets of the samples or species are required. This can be easily achieved, without the need to create large numbers of separate datasheets, by temporarily selecting subsets from a single sheet, analysing them (and thus creating new branches on the Explorer tree, with the results windows listing the selection used for any particular branch), and then restoring the full data set. There are several different ways to select subsets, described below, but it is important to keep in mind the distinction between highlighting and selection. The act of clicking on a row and/or column header *highlights* that row and/or column; it does not *select* it. Once you are happy that you have highlighted the correct set of samples (and/or variables) you can select them using the **Select>Highlighted** menu. Highlighting is just an intermediate stage, and has functions other than selection (e.g. to identify samples that need individual transformation, whilst the rest of the matrix remains unchanged - see next section). Alternatively, highlighting can be bypassed altogether and selection made by other direct choices from the **Select** menu.

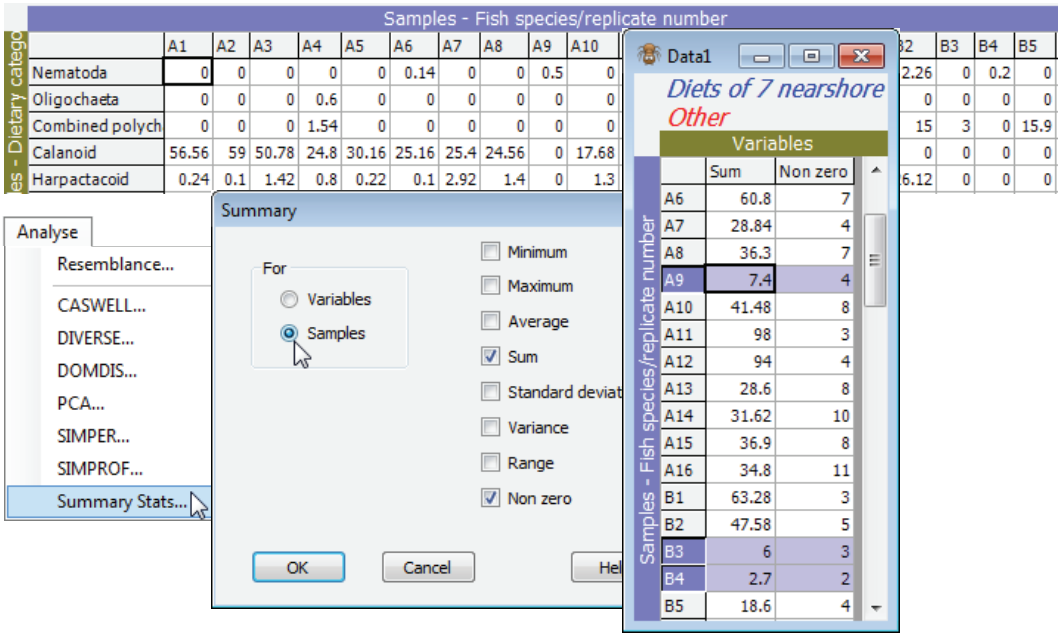
(W Australia fish diets)

Dietary data on the gut contents of 7 marine fish species found in nearshore waters of the lower west coast of Western Australia are reported by Hourston M, Platell ME, Valesini FJ, Potter IC 2004, *J Mar Biol Assoc UK* 84: 805-817 and Schafer LN, Platell ME, Valesini FJ, Potter IC 2002, *J Exp Mar Biol Ecol* 278: 67-92. Data is %volumetric gut contribution (reflecting both composition and gut fullness) of each of 39 'dietary categories' (broadly classified taxa), in a total of 68 samples across the 7 fish species (unbalanced replication), each sample being from a pool of 5 fish guts. The data matrix in PRIMER 7 format can be found in C:\Examples v7\WA fish diets\WA fish diets %vol.pri. Since species are involved in the definition of both samples and variables it is important to keep a clear head as to which are which! Here the fish predator species are the sampling device, so different individual fish guts (in pools of 5) constitute the samples. The assemblage studied is the set of prey species (higher taxa) making up the dietary categories; they are the variables.

Summary Statistics

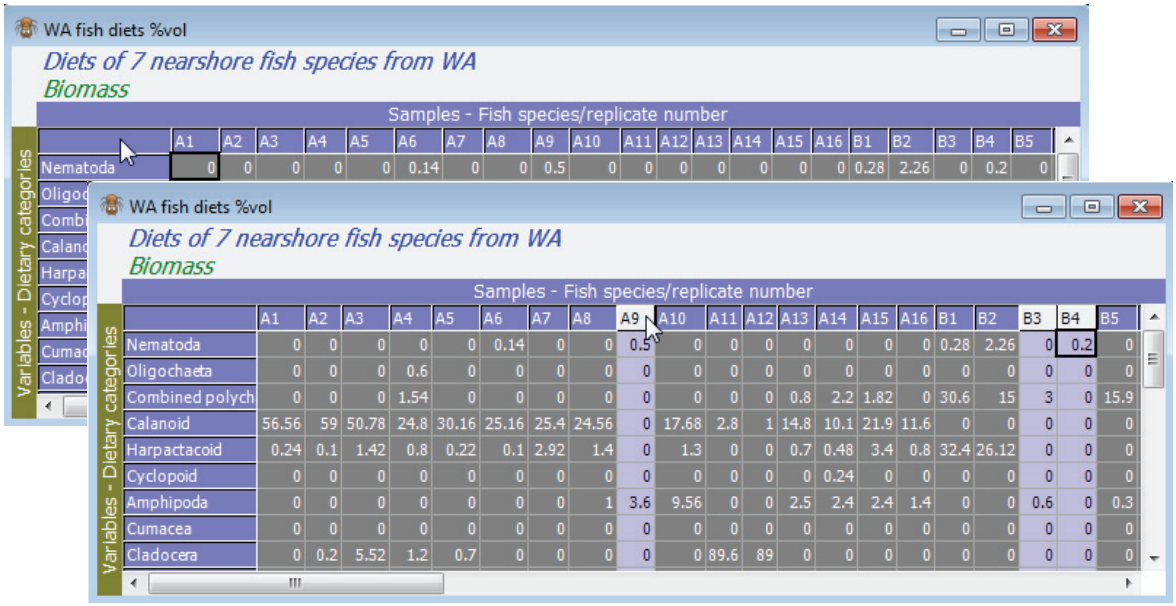
v7

File>Open>Filename: WA fish diets %vol, and examine the factors sheet with **Edit>Factors**. The samples form 7 groups (identified in the labels by A to G) which are the different predator species, three of which, B: *Sillago schomburgkii* (n = 10), E: *Sillago bassensis* (n = 14), G: *Sillago vittata* (n = 16), are from the same genus (congeneric) and thus of particular interest in terms of whether their diets are distinguishable (they occupy different niches in the 'dietary space'). First, calculate simple summary statistics for each sample with **Analyse>Summary Stats**>For•Samples. Not all summary options (Min, Max, Average, Sum, Standard deviation, Variance, Range, Non zero) may be meaningful in particular contexts: one that is informative here is ✓Sum. This shows that three samples (A9, B3 and B4) have low total gut fullness (<<10%), even though from a pool of 5 guts, and it is justifiable to look at the effect of (temporarily) dropping these samples from the analysis on the grounds that they contain little information on dietary composition (and could thus have large variability in similarity with other samples, see Section 5 on zero-adjusted Bray-Curtis).

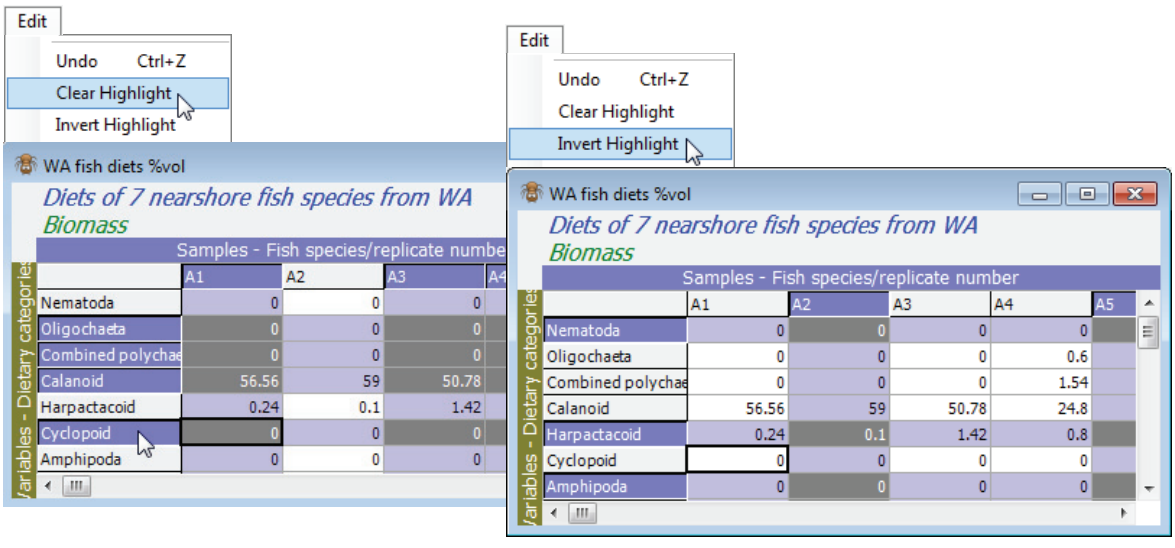


Control of highlighting

Thus, with the WA fish diets %vol datasheet as the active window, highlight all columns *except* the three samples A9, B3 and B4. There are various ways of doing this. Clicking on a column label highlights that column (in light blue shading if the default Windows colours are used) and is a toggle action (a second click turns off the highlighting). Clicking, holding and dragging the cursor across column headers will highlight a sequence of samples, as will the usual Windows action of clicking on the first, then holding down the Shift key when clicking on the last. (The Ctrl key has no effect; also the toggling action is set so that intermediate columns which are already highlighted will not be turned off if a wider range of columns, including them, are highlighted in these ways). However, the easiest way of highlighting all except a few columns is to highlight all the data, by clicking in the blank cell at the top left of the sheet, then click on the A9, B3 and B4 labels to de-highlight just those. (The top left cell is also a toggle note, so a second click is a convenient way of clearing all highlights, though this can also be done by **Edit>Clear Highlight**).

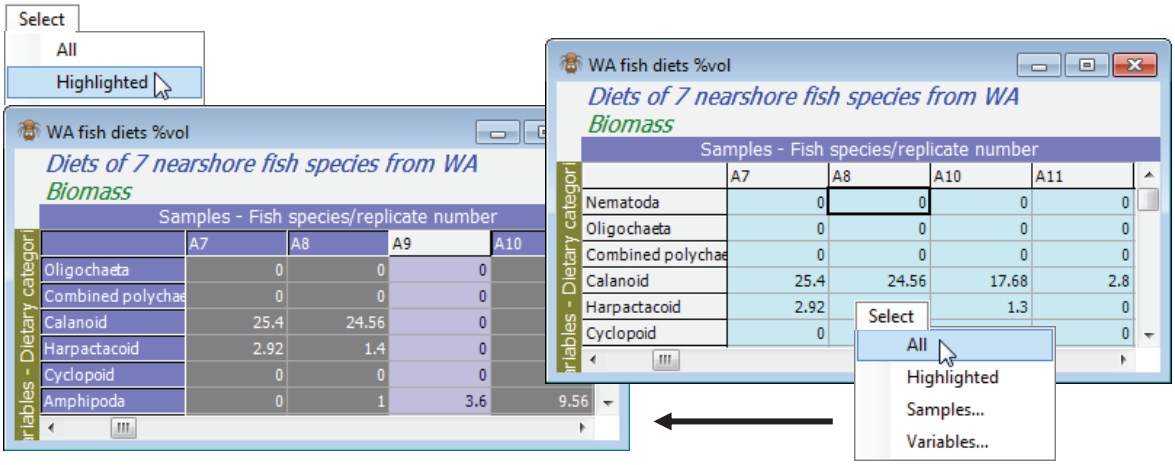


In the default Windows colours, cells in the table have one of three backgrounds: very light grey, light blue or dark grey. Three colours are necessary because highlighting can also be by rows, or rows *and* columns simultaneously. The rule is that the cells with the darkest background are those that are highlighted. You will see this best by turning off all highlights then clicking on a random set of row and column labels: the intersections are considered the highlighted part of the matrix. (Individual cells in the table cannot be highlighted by clicking on them; it is not meaningful to be able to select, say, only A1 Calanoids and B5 Amphipods. It is best not to think of the data as a conventional spreadsheet: only a limited set of operations make sense for sample × variable arrays). Note that highlights can also be inverted by **Edit>Invert Highlight**.



Selecting & deselecting highlights

When all except columns A9, B3 and B4 are highlighted, take **Select>Highlighted**. Alternatively, right click when over the data and a drop-down menu will appear, of operations from the **Edit** and **Select** menus, including **Select highlighted**. The matrix entries now have a different (turquoise) background indicating that you are operating with a selection – a new datasheet window is not created and the non-selected data is not lost. The operation can be simply reversed by deselecting the highlight with **Select>All** – the highlights are retained so it is easy to change some of them (or reverse them with **Edit>Invert Highlight**, see example above) and reselect.

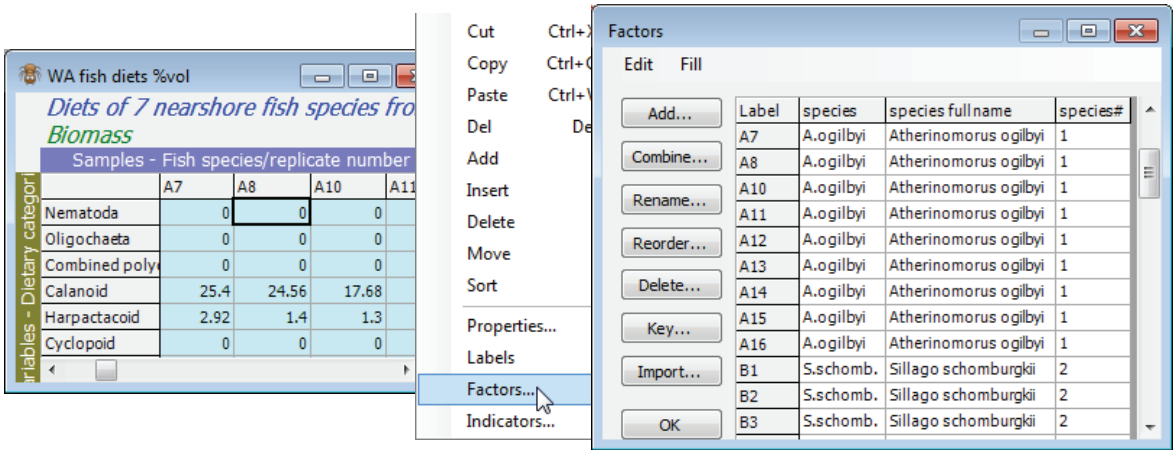


Duplicating a selected worksheet

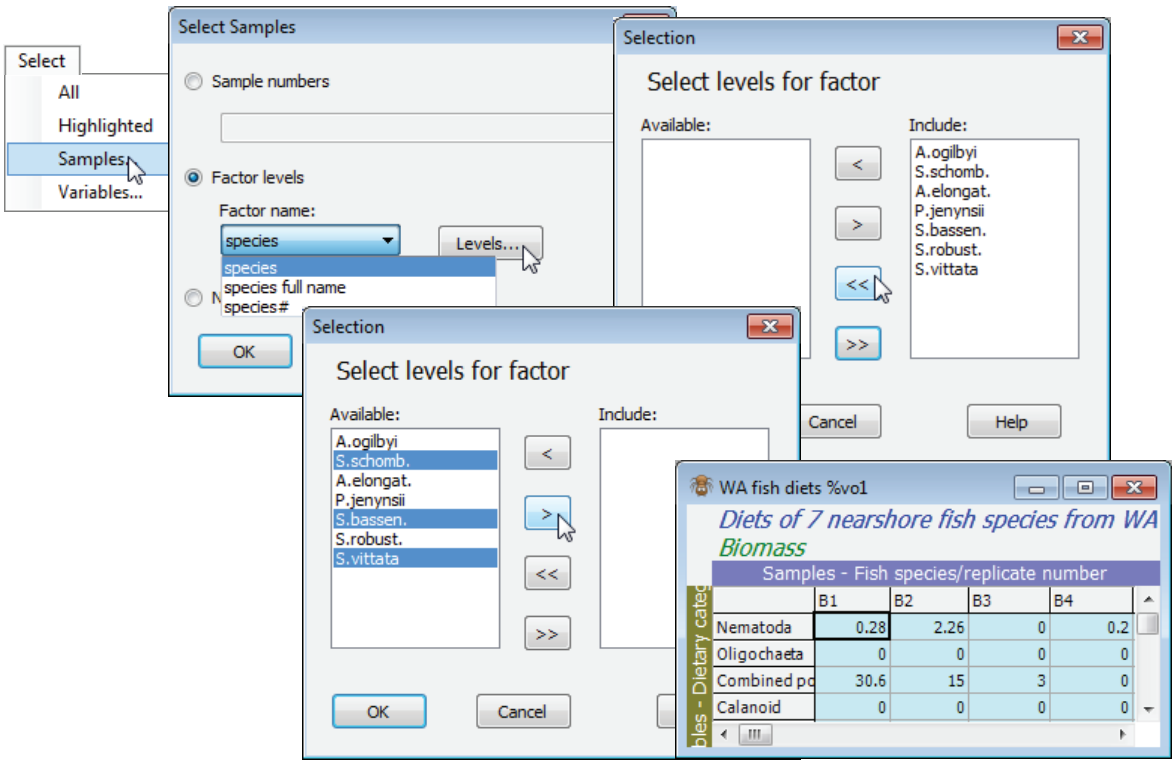
Though most Save operations are on whole workspaces, occasionally a data matrix needs to be saved externally, perhaps because it is needed in a different workspace or with other software. In order to protect against overwriting an original, external, data file with a version which is under a (possibly temporary) selection, **File>Save Data As** will ignore selections and save the whole data. To force a save of only the selection, you must first duplicate the selected sheet, with **Tools> Duplicate.** Do this on the selected form of the WA fish diets %vol data – which has excluded A9, B3, B4 – to create a new datasheet, Data2, which will now not contain these samples when saved.

Selecting by factor levels

The highlighting route to selection can be bypassed altogether using the other options on the **Select** main menu, **Select>Samples** and **Select>Variables** (and an example of the latter was seen in the previous section). Here, to select only those samples from the three congeneric *Sillago* predator fish species (labels starting B, E or G), it would be neater to use the factors that have already been set up to identify these different levels: S.schomb., S.bassen., S.vittata from the factor species, or the non-abbreviated species full name factor, or equally, 2, 5 and 7 from the numeric factor species#.



From the WA fish diets %vol datasheet, take **Select>Samples>•**Factor levels>(Factor name: species)>**Levels**, giving a standard Selection window, with boxes listing levels to Include, and those Available but not included. Move back all items to the Available list with <<, then using the > button move back the desired levels: S.schomb., S.bassen., S.vittata to the Include list. This can be either singly, or all of them can be highlighted with Ctrl clicks (a range would use Shift click), in the usual Windows manner, and then all taken across to the Include box with >



Multiple selections

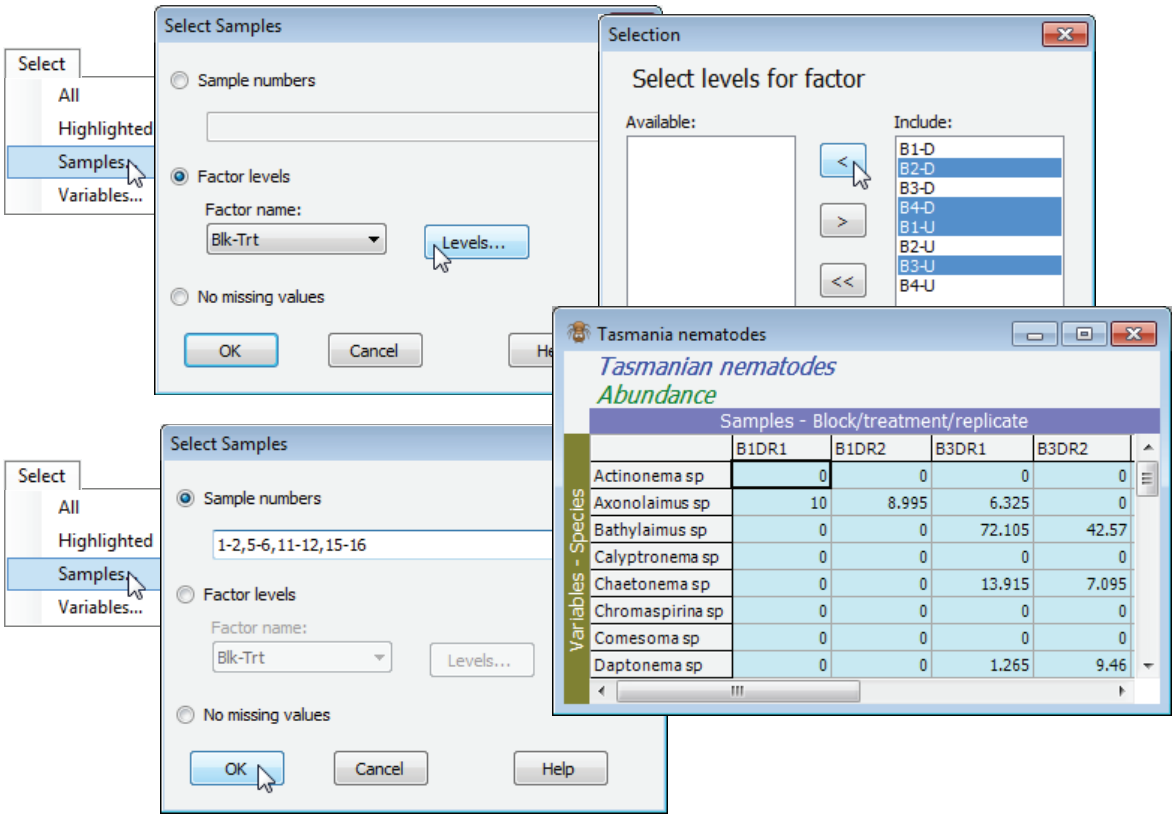
It is important to note the effect of this second selection on WA fish diets %vol; it produces a sheet of all samples from these three *Sillago* species. The prior exclusion of samples A9, B3 and B4 has been ignored – each new selection is a fresh operation on the full data array that is held in that worksheet. If, as seems likely, a compounding of the two selections was required, then that is easily achieved, in at least two ways. One would be to take the current selection (all B, E and G samples), highlight everything displayed (e.g. by clicking in the blank, left corner box), dehighlight B3 and B4, by clicking on their column labels, then **Select>Highlighted**. (This is logically sound because all the omitted A, C, D, F samples from the first selection are not highlighted at that point.) This would retain the single copy of WA fish diets %vol in the Explorer tree. A more general option though, which would be more relevant to a complex multiple sequence of selections, is simply to **Tools>Duplicate** the sheet after every selection, then do the next selection on the new sheet. So, if the above selection of the *Sillago* samples had taken place on Data2, samples B3 and B4 would automatically have been excluded. Note, however, that if the two selections are on different axes (selecting a subset of both samples and variables) then they will not interfere with each other, i.e. when sequentially taking **Select>Samples**>•Factor levels and **Select>Variables**>•Indicator levels.

A third option for repeated selection of samples, with the outcome of multiple selection being a single worksheet (rather than a series of copies), is to create a compound factor (with **Factors> Combine**), which will allow simple selection of one (or more) of its levels.

To illustrate this, save and close down the above workspace, as WA fish ws(.pwk), and re-open the previous workspace, C:\Examples v7\Tasmania meiofauna\ Tasmania ws. Here there are only 16 samples, which helps for illustrative purposes (though in the real context would make selections quickest by simple highlighting). The study design has two crossed factors: *Trt* (disturbed, D, and undisturbed, U, sediment patches), and *Blk* (4 areas of sand-flat, B1 to B4), with 2 replicates in each combination. An example of 2-factor selection for the Tasmania nematodes sheet would be to select distinct sand patches within each treatment, say blocks 1 and 3 for D, and blocks 2 and 4 for U (which would make the data sheet 2-factor nested rather than crossed). Use the *Blk-Trt* combined factor created in the previous section to **Select>Samples>**•Factor levels>(Factor name: Blk-Trt)> **Levels**, leaving B1-D, B3-D, B2-U, B4-U in Include and moving the others back to Available.

Selecting by number and non-missing

It may sometimes be easier to use the sample numbers, here **Select>Samples>**•Sample numbers> 1,2,5,6,11,12,15,16, though this is more likely to be useful where such numerical lists are output in results (e.g. by the BEST routine, Section 13), and can be copied and pasted into this dialog box.



The final possibility is **Select>Samples>**(•No missing values) in which only those samples which have no entries of Missing! for any of their variables will be selected. Missing! entries are unlikely for species matrices (as here) but this facility might be useful sometimes for environmental arrays, to find samples which have a complete set of variables.

Selecting variables

Any of the options for selecting samples are also available for selecting variables, e.g. selecting by variable numbers or by levels of an indicator, the latter as seen in the example of the previous section, in which the Tasmanian copepods of 'Undetermined taxa' were excluded. There is a similar construction of selecting variables with no Missing! entries across the full set of samples. Note that if the selection option of (•No missing values) is chosen for both samples and variables, the order in which these are taken will affect the outcome. In practice, if it is required to form a complete matrix (and this is now less essential than in previous versions of PRIMER since all resemblance measures are now defined under pairwise-elimination of missing values, Section 5), a more careful manual deselection of the array rows and columns is likely to be preferable, utilising knowledge of which are the most important samples or variables to attempt to retain. Alternatively, where the data can be approximated by multivariate normality, missing entries can sometimes be successfully estimated by the EM algorithm – see the **Tools>Missing** menu, in Section 12.

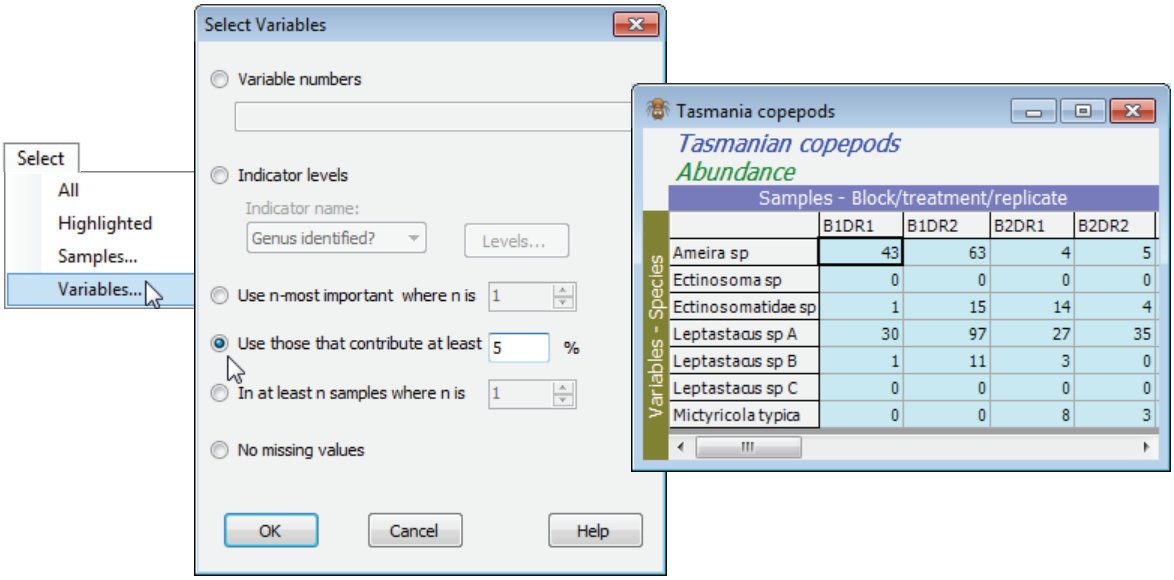
v7

Selecting by 'most important'

There are, however, three other selection methods under **Select>Variables** that are specific to selecting species (or other taxon-type) variables, in which matrix entries are positive 'amounts' of that species (counts, biomass, area cover etc). The idea of the first two options is to be able to drop species which are not a substantial component of the overall counts (or biomass, area cover etc) in any sample. The third option, an addition to PRIMER 7, is to drop species which occur in fewer than a specified number of samples, e.g. **Select>Variables>**(•In at least *n* samples where *n* is 2) would drop species which were only seen on one occasion. (It is important to note, however, that removing low abundance or rare species in this way is not required for most of the methods in PRIMER, based on Bray-Curtis similarities for example, and should be done only where there is good reason, e.g. when using a resemblance coefficient which is sensitive to rare species – such as chi-squared distance or Gower, Section 5). The option to **Select>Variables>**(•Use those that contribute at least 5 %) applied to the copepod counts in Tasmania copepods would drop species which, for every sample, account for <5% of its total abundance, leaving only 7 of the original 17 species in the selected sheet. Alternatively, the number of species to retain can be specified, rather

v7

than the %, but the principle is the same. Taking **Select>Variables>**(•Use *n*-most important where *n* is 7) generates the same set of species, naturally. If *n* is larger, say 10, then to be retained, the threshold percentage that a species must contribute somewhere will drop – in fact a threshold of around 3% will leave 10 species. If *n* is smaller, say 5, then a higher percentage cut-off is needed (10% in fact). The algorithm simply varies the cut-off percentage until the matrix retains only the exact number of species *n* requested. This means of selecting 'important' species (rather than by taking their total abundance across all samples and selecting the top *n*-ranked of those) is preferable because it retains species which are important in impoverished sites, with low total abundance.



The point is re-iterated that **Select>Variables** will operate in combination with **Select>Samples** (unlike repeated **Select>Samples** or **Select>Variables** operations on their own), to ensure the behaviour that would be expected. That is, if a sample selection is in operation then the 'most important' 10 species – or the species which occur in at least 2 samples – are determined only with regard to that selection, not using all the samples.

Close the Tasmania ws – there is no need to resave it, since when met in a later section it will not be for a subset of either the samples or species.

Selection in resemblance matrices

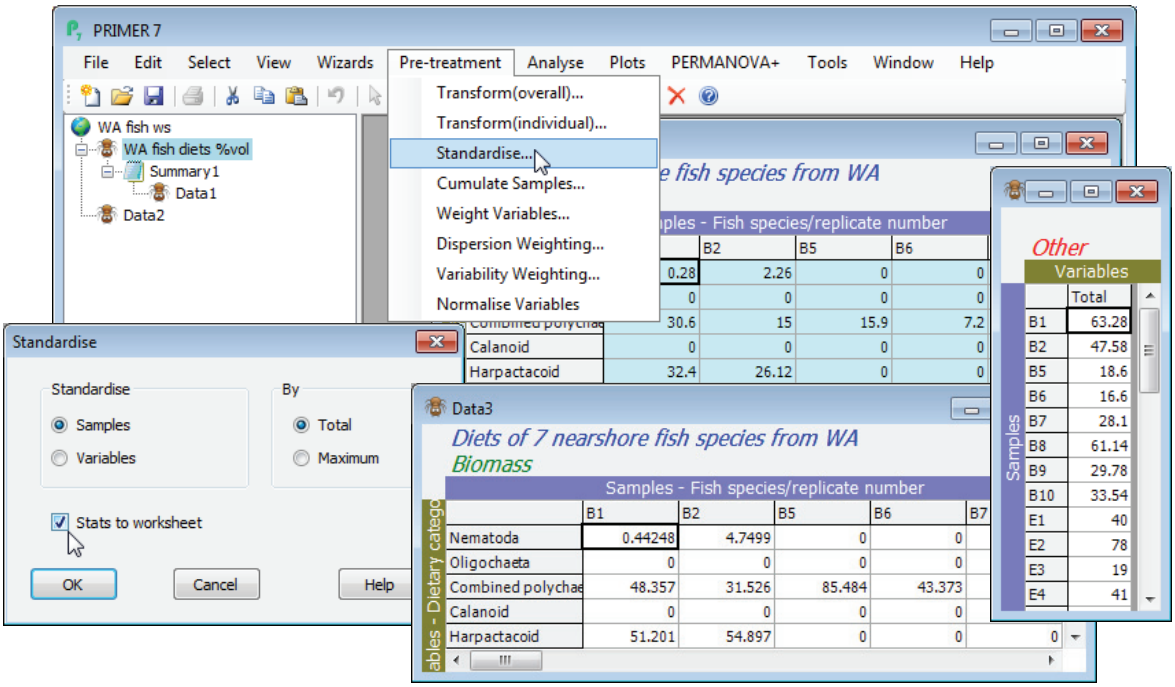
Looking ahead to Section 5, when the active window is a (triangular) resemblance matrix, selection can take place just as for a (rectangular) datasheet, by **Select>Highlighted** or **Select>Samples>** (•Sample numbers) or (•Factor levels). Another option is provided in that case: selection of only the rows and columns containing at least one value above or below a specified threshold by, for example, **Select>Samples>**(•Values>0.95), or selecting only rows and columns containing at least one Undefined! resemblance entry, by **Select>Samples>**(•Undefined values). These are mainly used for picking out, in the first case, collinear environmental variables from a large correlation matrix (values > 0.95 or <–0.95 say), Section 13. In the second case, this might more easily identify similarities that are undefined because neither sample contains any species at all, in cases where the similarity measure (such as Bray-Curtis) treats such samples as uninformative, Section 5.

4. *Pre-treatment* options

Standardising samples

How the data are treated, prior to computation of a resemblance matrix (e.g. similarities), can have an important influence on the final analysis, and such decisions often depend on the practical context rather than any statistical considerations. For example, standardising the samples (by total) divides each entry in the data sheet by the total abundance in that sample, across all variables (species). This would turn assemblage counts for each sample into relative percentages (what is referred to by statisticians as compositional data), all samples then adding to 100% across species. It thus removes all differences in total abundance in each sample from the multivariate comparison of samples. Sometimes this may be desirable, e.g. where the unit of sampling cannot be tightly controlled. An example is the data we have just been working with (on W Australian fish diets), analysing the prey taxa in the gut contents of fish predators: the quantity of food in the gut varies across individual fish in an uncontrollable way so is not relevant to a multivariate comparison of the prey composition, and the data should initially be sample-standardised. On the other hand, a typical marine impact study, using sediment-dwelling fauna sampled by a corer of fixed size, more strictly controls the quantity of material in each sample. It might then be important to use the fact that a potentially impacted site contains 5 times fewer individuals, in total, than a control site, so sample standardisation would be undesirable. The philosophy in PRIMER 7 is that users control all such pre-treatment decisions, combining them in an order under their choice, appropriate to the context. Each pre-treatment step results in display of a revised datasheet so the user can see its effect, before proceeding to analysis (or in some cases a further pre-treatment step).

Re-open the workspace WA fish ws from the directory C:\Examples v7\WA fish diets, or if not previously saved, **File>Open**>Filename: WA fish diets %vol.pri. (Note that if you had a selection in place at the time the workspace was saved, this will still be operational. You can leave this on or deselect it with **Select>All**, but it might make sense to leave samples A9, B3 and B4 excluded, because of their very low sample totals – gut fullness <<10% – and thus unreliable % composition after standardising). Take **Pre-treatment>Standardise>**(Standardise•Samples) & (By•Total) & (✓Stats to worksheet). You will see from the resulting sheet (probably named Data3) that samples are now expressed as % composition of each prey category, the columns adding to 100.



Stats to worksheet

Several of the routines in PRIMER 7 also incorporate a *check box* for sending summary statistics used in that routine to a further worksheet. Here, this results in a second sheet (probably named Data4), which is just a single column of totals across prey species for each of the gut samples. This is the same data as previously obtained with **Analyse>Summary Stats>**(For•Samples) & (✓Sum), in Section 3. (There is often more than one way of obtaining the same information in PRIMER!).

Another example of summary statistics being sent to a separate worksheet is for the Normalise pre-treatment option – see below – for which the mean and standard deviation of each variable, used in the normalisation process, can be sent to a separate sheet. If this option is not selected, the same information is usually sent to a text-format results window, which can be viewed from the Explorer tree but cannot be further manipulated (unless then saved as an external .txt or .rtf file and edited in a text editor – or directly copied and pasted into Excel column(s) – to re-input as a new sheet).

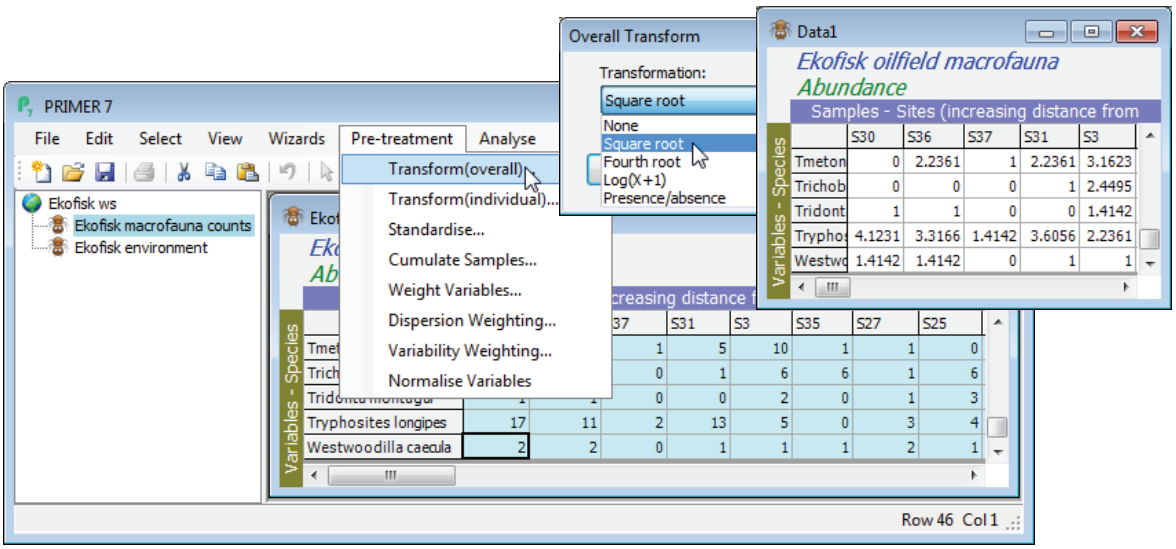
Standardising species

Pre-treatment>Standardise can also be used to standardise the matrix on the variables axis, e.g. to ensure that each species is given equal weight in any ensuing similarity calculation by making their totals across samples all add to 100, with (Standardise•Variables) & (By•Total). There is an alternative option, (By•Maximum), to scale each species so that its maximum value across samples is always 100. Similar species standardisations are already built into some resemblance measures, e.g. the Gower coefficient S_{15} (see Section 5) scales all species to have the same maximum, in effect, because it divides each variable through by its range (and for most species the minimum value across all samples is usually 0). For analysis of community samples, however, such species standardisation (by totals or maxima) is usually undesirable because it gives rare and very low abundance species as much weight (usually more weight in practice) than common and abundant ones. Variable standardisation, over samples, occasionally has a role with non-assemblage data which is still in the form of positive 'quantities', taking values down to zero but on non-comparable scales. Generally more useful though, for environmental-type matrices (where measurement scales differ, zero may play no special role at all, and values can be negative, especially after transformation) is what PRIMER refers to as normalisation – removing both scale and location differences amongst the variables (see below). The major use of variable standardisation is for the multivariate analysis of species rather than samples. To avoid the problems of standardising rare species, this first requires reduction to the 'most important' species, using the techniques of the last section. Then species standardisation is an important step in determining groups of species that display a *coherent* response across the set of samples; see Section 10 and Chapter 7 of the CiMC manual. **File>Save Workspace** the current form of the WA fish ws, for further use later, and close it.

Transforming (overall)

Transformation is usually applied to all the entries in an assemblage matrix of counts, biomass, % area cover etc, in order to downweight the contributions of quantitatively dominant species to the similarities calculated between samples (see Chapters 2 and 9 of CiMC). This is important for the most commonly-used resemblance measures like Bray-Curtis similarity, which do not incorporate any scaling of each species by its total or maximum across all samples The more severe the initial transformation, the more notice is taken of the less-abundant species in the matrix. It is for the user to choose a balance between contributions of dominant and less abundant species, in the specific context, by picking from the sequence: None, Square root, Fourth root, Log(X+1) and Presence/absence. (Reduction to presence/absence, i.e. 1/0, is thought of as a transformation since it would be the logical end-point of taking ever more severe power transforms: square root, 4th root, 8th root, …, and it is clearly one way in which less abundant species are given a similar weight to abundant ones.) If standardisation of samples by total is also required, for example to ameliorate the effects of differing sample volumes, it is logical to standardise first, then transform.

Open the previously saved workspace Ekofisk ws from the C:\Examples v7\Ekofisk macrofauna directory (or open Ekofisk macrofauna counts.pri and Ekofisk environmental.xls into a clear workspace, Section 1). **Edit>Properties** on the count matrix shows that 173 species were found across the 39 sites, which are ordered in increasing distance away from the oil-field centre (the putative source of a pollution gradient, diluting with distance). It is crucial to stress at this point that an initial reduction in the number of species entered into the later multivariate analyses of these samples is not required - as just remarked, it is the job of the transformation and the similarity measure to balance contributions from abundant and rarer species. However, purely in order to visualise the effect of the differing transformations, on a more manageable number of species, take **Select>Variables**>(•Use those that contribute at least 2 %), which selects 46 'most important' species (you can see that is it 46 by clicking in the last row of the selected array, when the row and column position of the cursor will be seen at the bottom right of the PRIMER desktop). On this reduced matrix, take **Pre-treatment>Transform (overall)>**(Transformation:Square root), and also the options for Fourth root and Presence/absence. Rename the four 'Data*n*' sheets appropriately, e.g. by clicking twice (slowly) on their name in the Explorer tree and typing in Square root etc.

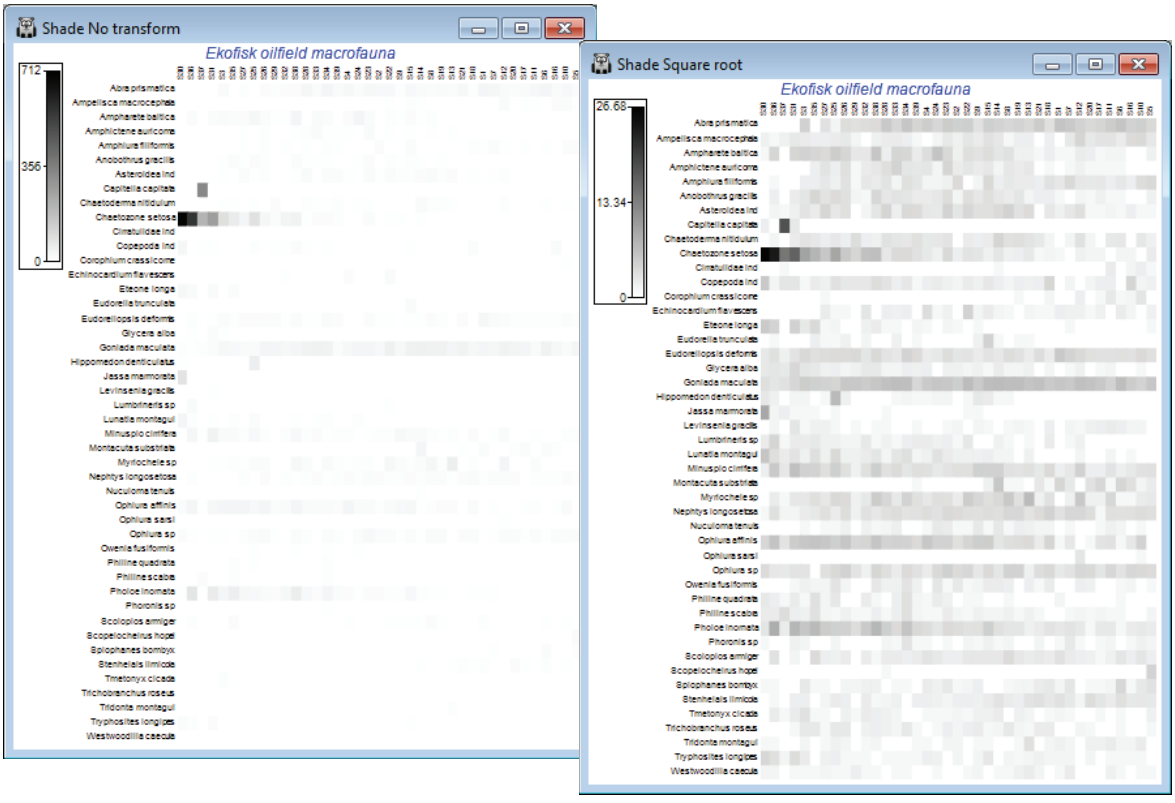


Shade plots to aid choice of transform

v7

A major new feature in PRIMER 7 is the large number of additional plotting routines, one of the conceptually simplest but most powerful being Shade Plots, which are simple visualisations of the data matrix, with darker (or different colour) shades in each cell of the array representing higher abundances. White space denotes the absence of that species (row) in that sample (column) and full black the maximum abundance (or biomass etc) in the array. Grey (or one/two colour) shades are linearly proportional to the intermediate abundances, as shown in a shade/colour key. Clarke KR, Tweedley JR, Valesini FJ 2014, *J Mar Biol Assoc UK* 94: 1-16 demonstrate the usefulness of shade plots in getting a 'feel' for a sensible choice of transformation for the context, e.g. if an assemblage analysis needs to take account of a wide range of common and less abundant species but the current shade plot is largely a sea of white space – because at the current transformation most abundances are still dwarfed by those for the dominant species – then the need for a heavier transformation is immediately seen. At the opposite extreme, if most of the cells from species which are present are displayed at about the same (dark) intensity then the data is likely to have been overtransformed into, effectively, presence or absence, and this may not be the required quantitative analysis.

On both the original and transformed Ekofisk macrofauna sheets take **Plots>Shade Plot**, to give:



v7

The choice looks to be between square root and fourth root, but note how the fourth-root matrix largely reflects the P/A structure, with the quantitative information little used. And after restoration of the 125 species (<2% of the composition anywhere and temporarily eliminated, purely for clarity of the plots here), they are also likely to add a great deal of random 'noise' on this scale. At the other extreme, the previous page shows that a failure to transform at all would leave a multivariate analysis (based on a measure such as Bray-Curtis) dependent only on a small handful of dominant species. Be aware of the dangers of 'choosing the transformation which gives you the answer you want!' but these plots suggest that the (relatively mild) square root transform might be relevant for data of this type (macrobenthic studies around N Sea oil-fields) – allowing the abundant species to play a greater role, but also taking into account contributions from a wide range of less-dominant species. Whether a multivariate analysis can discern any pattern of change with distance from the oil-field is more open to question, on the basis of this plot! The sites (*x* axis) are ordered from left to right in increasing distance from the oil-field but a matching trend in assemblage pattern is quite hard to discern (but is clearly present – see Section 8). We shall see later that astute re-ordering of the *y* axis (species) is visually helpful here (though a multivariate analysis ignores the ordering of variables!), and can be accessed from the **Graph>Special>Re-order** menu. Discussion of the wide range of possibilities on this dialog is deferred until Section 10, under **Wizard>Matrix display**.

Transforming abiotic variables

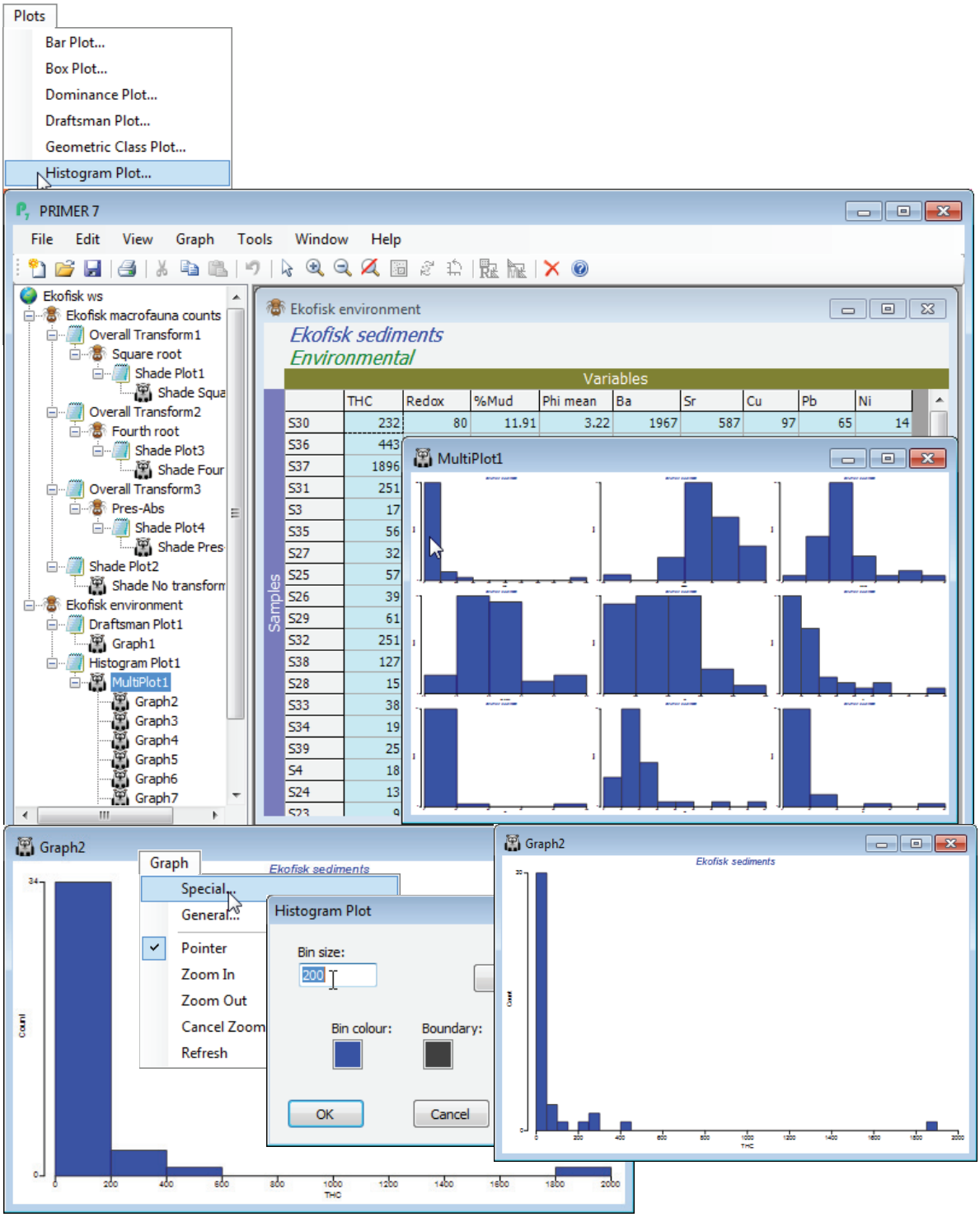
Transformations may be appropriate for environmental variables too, though usually for a different reason (e.g. in order to justify using Euclidean distance as a dissimilarity measure on normalised variables). However, these are usually selective transformations, required only for some variables, and with different transforms potentially applicable to variables of different types. The global **Pre-treatment>Transform (overall)** applies the same simple power or log transform to all variables, whereas **Pre-treatment>Transform (individual)** operates only on highlighted portions (usually sets of variables), and can allow user-defined expressions if a specific formula is appropriate to a certain variable. More sophisticated data manipulations with user-defined expressions are deferred to Section 11; here we concentrate on one or two commonly used transforms for abiotic variables.

In the Ekofisk ws, the Ekofisk environmental sheet holds 9 variables measured on sediments at the same set of 39 sites: total hydrocarbons (THC), several heavy metals, redox and two particle size measures, % mud and ϕ mean. The first variable in the sheet is just distance in km from the oil-field centre, not of itself a measure which organisms in the assemblage will respond to, and which should not be used for any assessment of the pattern of environmental change around the field.

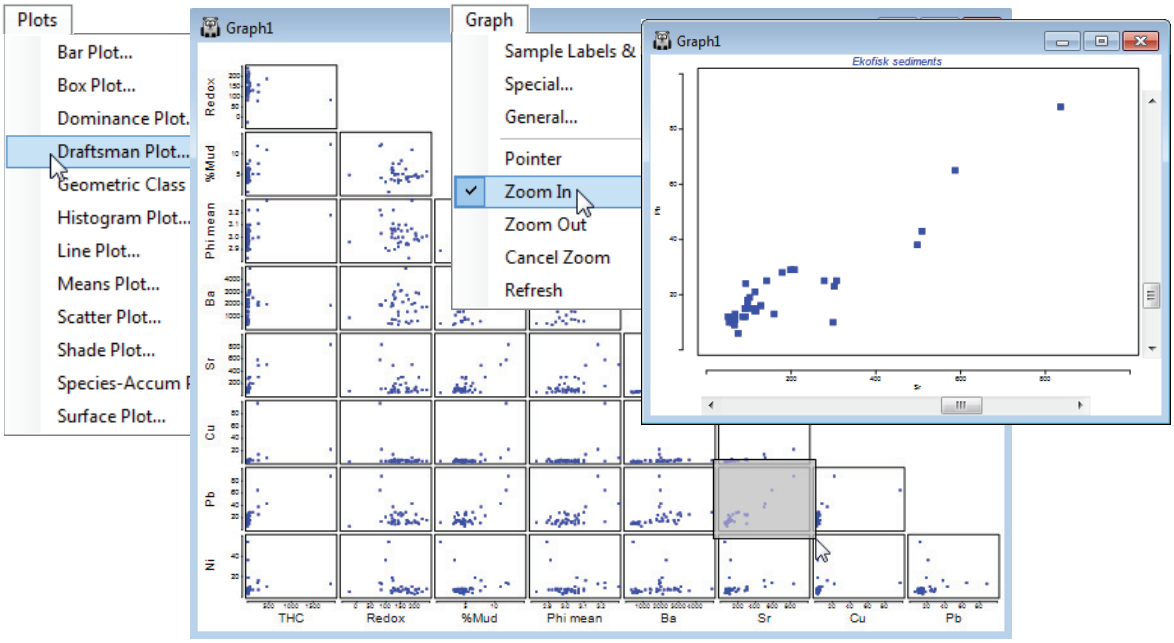
Draftsman, histogram & multi-plots

v7

Temporarily deselect the Distance (as in Section 3), and run **Plots>Draftsman Plot** on the other 9 variables; also **Plots>Histogram Plot** (a new plotting feature in PRIMER 7). The latter leads to an example of another new feature, a *Multi-plot* (see Section 7), in which the window is divided into row and column cells (the numbers of which are under user control), each of which contains a standard graphics window. The multi-plot can hold graphs of different types (e.g. a multi-plot which will often be met mixes MDS ordinations and their associated Shepard diagrams, Section 8) but typically all component plots are of the same type, as here when they hold histograms for each of the 9 variables. Clicking on a cell of the multi-plot will cause that component plot to be shown at normal size and able to be operated on, in terms of changing axis labelling, titles etc. These general editing operations for plots are covered in Section 6, but each plotting routine has some specialised operations that apply only to that plot type and the one that might be of use here is to change the histogram bin width, e.g. for the THC histogram take **Graph>Special**>(Bin size:50). To shut down the full version of the component plot simply close it with and click on a further component.



The Draftsman plot is simply a set of pairwise scatter plots of all 36 combinations of the 9 variables laid out in a single lower triangular graph array (this is not a multi-plot note, though individual portions of the plot – down to single scatter plots – can be viewed by *zooming* into the plot, a general feature available with all graphics through the icon or the **Graph** menu, Section 6). Whilst histograms would often be used to look at the distribution of individual variables over the samples, the scatter plots of the draftsman plot can be an equally effective way of viewing this, especially if there are too few samples to bin into a meaningful histogram.



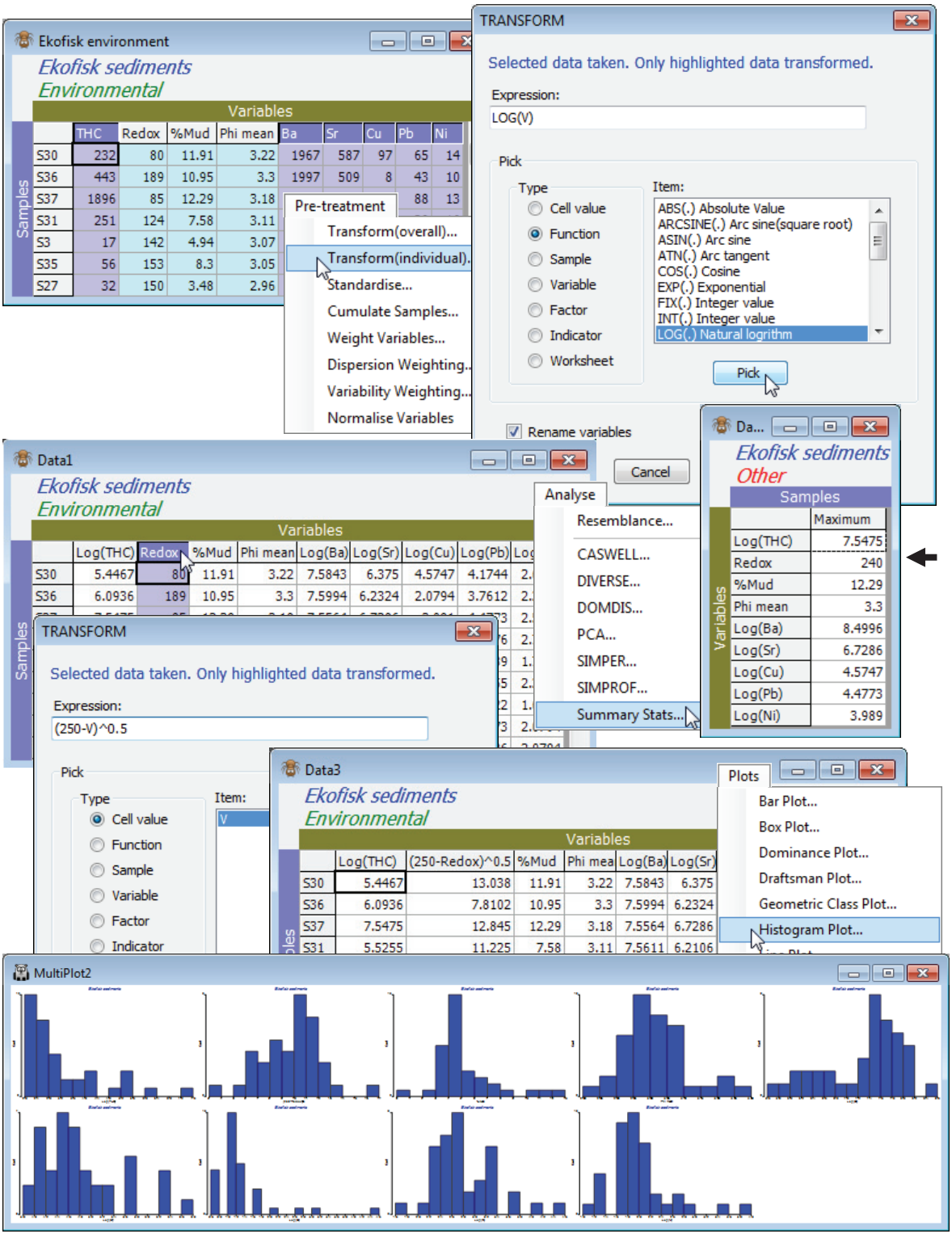
Transforming (individual)

Both the Draftsman and Histogram Plots show that several of the Ekofisk abiotic variables are highly right-skewed (tail to the right), and it would be wise, if we are to limit the distorting effects of outliers and normalise the data to a desired common measurement scale, to subject THC and the heavy metal concentrations to a strong transformation such as $\log(x)$. The particle size variables do not need further transformation (ϕ mean is already on a log scale). There is a case for regarding Redox as left skewed (it certainly has a large negative outlier), so we shall take the opportunity to demonstrate how to achieve a (mild) reverse power transform: $(a - x)^b$.

Highlight the THC, Ba, Sr, Cu, Pb, Ni variables and take **Pre-treatment>Transform (individual)**. The transform operation itself can be any of the Transform(overall) options: square root, fourth root, log, reduction to pres/abs, using the Expressions: Sqr(V), V^0.25 (≡ Sqr(sqr(V)), log(1+V), PA(V) respectively, in which V (value) stands for any highlighted data entry (note that upper or lower case is not important in the expressions). But it is not limited to these: many other transforms can be constructed. In fact any expression using the Basic language syntax is permitted, involving operators: +, -, * (times), / (divide), ^ (power); functions: Sqr, Log (to base e) etc as above, and Abs (absolute value), Atn (arctan), Exp (exponential), Int (integer part of a number) and many others; and even logical operators: =, <, >, <=, >=, which return –1 if true, 0 if false. (An example of the latter might be to draw attention to cells with large counts using an expression like V>1000). For a comprehensive list of expression options take **Help** on the Transform dialog box and click on Transform expression. Operations can extend still further, to generate new entries as combinations of samples or variables (and even factors or indicators or other worksheets), but examples of these are deferred until Section 11. In this case you simply need the Expression: log(V) which you can type directly into the Expression box or select the function from the Pick box: (Type•Function) & (Item: Log(.) Natural logarithm)>**Pick**. The action of the **Pick** button is to place the selected function around the default entry already in the Expression box (of just V). Check the expression is the one you intended and **OK**, to obtain a new sheet in which the concentration variables have been log transformed – their labels indicate this if you have left on the default of (✓Rename variables).

Note that the remaining variables have also been carried across to the new sheet but untransformed. This is the result of only highlighting the requisite variables rather than fully selecting them, with

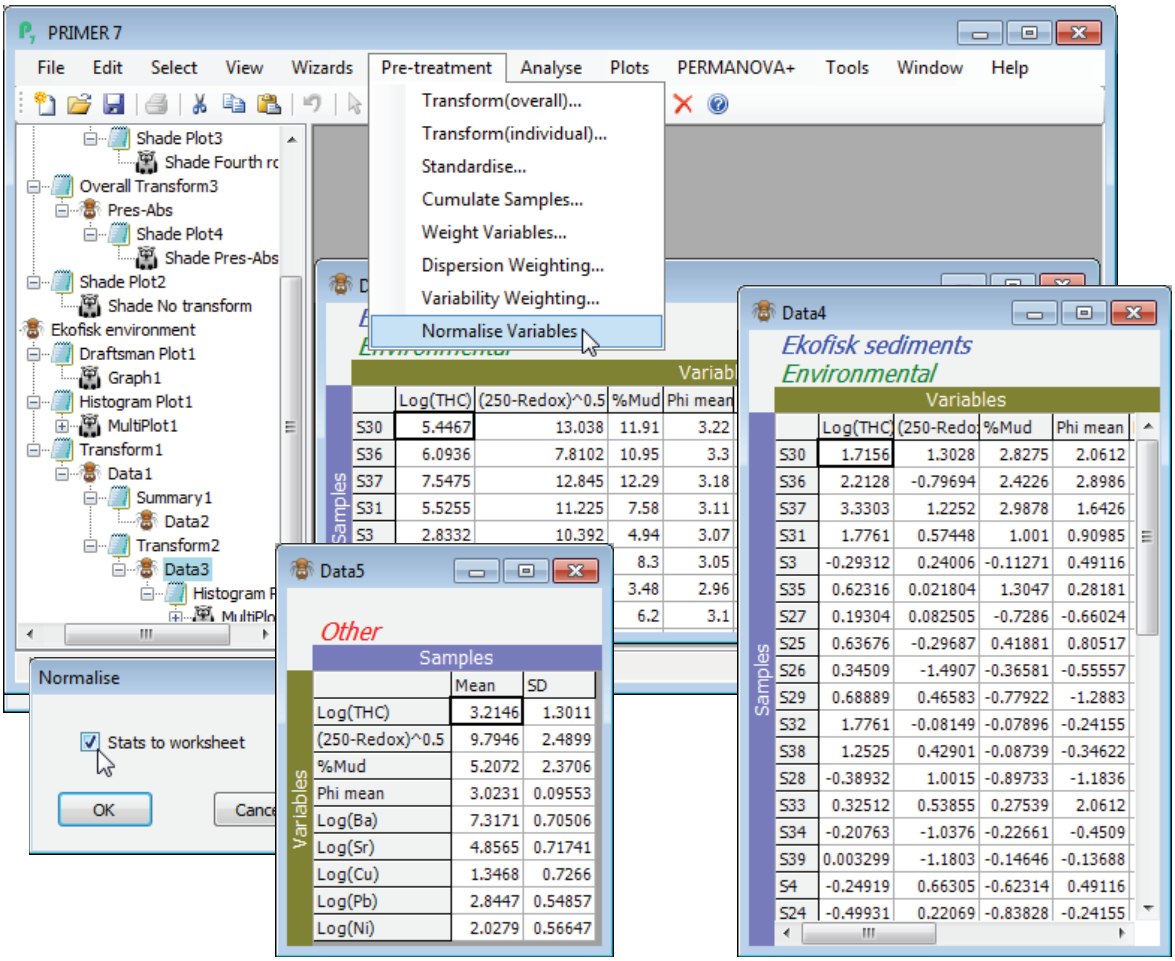
Select>Highlighted. (Had you done the latter then only the transformed variables would have been carried across to the new sheet, and you would have had to Select the others from the original sheet and **Tools>Merge**(d) them with the new, transformed variables). Now highlight the single Redox variable on the new sheet and **Pre-treatment>Transform(individual)**>(Expression: (250-V)^0.5). This reverses the distribution around a value just larger than its maximum, turning a mildly left-skewed shape into a mildly right-skewed one, and then the square root transformation will tend to remove that (mild) right-skewness – stronger would be to use log(250-V). Finding the maximum value for a variable is now easy with **Analyse>Summary Stats**>(For•Variables) & (✓Maximum). Again a new sheet is produced with the required mix of log, reverse square root and no transforms on different variables, and the efficacy of these in reducing the effects of outliers can be seen by another set of **Plots>Draftsman Plot** or **Plots>Histogram Plot**.



Normalising variables

It is typical of a suite of physico-chemical variables (or biomarkers, water-quality indices etc) that they are not on comparable measurement scales, unlike assemblage abundances. All multivariate analysis methods are based on resemblances between samples that add up contributions across the variables. This make no sense if there is not a common scale (transformation does not help in this regard). If the similarity or distance coefficient does not have some form of internal adjustment to put variables onto a common scale (the commonly used Euclidean or Manhattan distance measures do not), then it is important to pre-treat the data to achieve this. The standard means of doing so is *normalising*. Literature terminology is inconsistent here, but what PRIMER means by normalising is that from each entry of a single variable we subtract the mean (across all samples) and divide by the standard deviation of that variable. This is carried out separately for each variable. It is simply a scale and location change, and does not change the shape of the histograms above, for example. It does not therefore 'convert the variable to normality' – this is essentially what the transformation is trying (approximately) to achieve – but it makes the mean 0 and standard deviation 1, so that all variables now take values over roughly the same limits: typically (for a normal distribution) the range –2 to +2 covers roughly 95% of the entries, making contributions to (say) Euclidean distance from different variables comparable, and effectively giving each variable the same weight. This process is sometimes known, especially in the statistical literature, as standardisation, but PRIMER reserves the term *standardise* for scaling positive quantities only, by dividing by their total or maximum. Standardisation would therefore not succeed in putting onto a common scale variables for which zero is not a meaningful (and attained) end point of the scale, as is true for of many abiotic variables, such as temperature. And in a marine context, salinities may fluctuate over a narrow – but still potentially important – range well away from zero; standardisation (of variables) would then be completely ineffective. Note that, unlike standardising, normalising only makes sense – and is therefore only offered – for variables, not for samples.

On the transformed environmental variable matrix from the previous page, take **Pre-treatment> Normalise variables**>(✓Stats to worksheet), and note how the resulting variables now take values over comparable ranges, roughly –2 to +2. They are now ready for entry to **Analyse>Resemblance** >(Measure•Euclidean distance), using the methods of Section 5. Save Ekofisk wk.



Dispersion weighting of species

When variables are on different measurement scales, there is little viable alternative to normalising each variable (as above) thus equalising, in effect, their contributions to the multivariate analysis. When variables are (ostensibly) on the same scale, e.g. species abundances, then their respective contributions to commonly-used similarity coefficients, such as Bray-Curtis, will differ, based on the relative magnitude of counts (or transformed counts). Larger abundances are always given more weight (unless 'transformed out' to purely presence/absence). This may not always be desirable, however. For example, some numerically dominant species may give highly erratic counts over replicate samples within a site (or time or condition), perhaps due to an innately high degree of spatial clumping of individuals (individuals of that species arrive in the sample in clusters). This is likely to add 'noise' rather than 'signal' to the multivariate analysis, and downweighting of such species is called for, in relation to other species which are not spatially clustered, but have the lower variance associated with Poisson counts (the individuals arrive in the sample independently of each other). The weighting is achieved by the **Pre-treatment>Dispersion weighting** procedure, (Clarke KR, Chapman MG, Somerfield PJ, Needham HR, 2006, *Mar Ecol Prog Ser* 320: 11-27), covered in detail in Chapter 9 of the CiMC manual.

The differential downweighting is carried out by dividing the counts for each species by their index of dispersion $\overline{D}$ (variance to mean ratio, a 'clumping' measure), calculated from replicates within a group (site/time/treatment etc), and then averaged across groups. The weighting is valid under rather general conditions, not unrealistic, but the original derivation did require: a) data to be real species counts, not densities standardised to some unit volume or substrate area; b) independent replicates within each of a set of sample groups, so that there is a basis for assessing within-group variance structure; and c) those replicates to be of a uniform size (strictly 'quantitative sampling'). Downweighting is only applied where a species shows significant evidence of clumping, this being tested by an exact permutation test, valid for the very small counts that are typical of many species. The resulting dispersion-weighted matrix has a common (Poisson-like) variance structure across species but unchanged relative responses of species in different groups. This is an important point: there is no attempt here to place greater emphasis on those species which best show up a given group structure (e.g. best separate control from polluted conditions). Such 'constrained' methods run the risk of circular arguments: selecting out only those species that tell you the answer you wanted in the first place! All that dispersion weighting does is divide through each row of the matrix (species) by a constant, so that a different balance of species contributions will be obtained by the subsequent analysis. These weights are calculated solely using information from replicates within each group, not across groups, so a consistent species (low variance-to-mean ratio within groups) will be given a high relative weight even if it shows no difference at all between groups.

If dispersion weighing of a count matrix is contemplated, this pre-treatment step must be carried out before any transformation. It may still make sense then to transform the dispersion weighted data sheet: a species which has large mean abundance at some sites, and is found in very consistent numbers in all replicates from those sites, will still tend to dominate the similarities. Transforming now has the strict objective of balancing contributions of consistent abundant species with equally consistent but less numerous species. Previously, it was really used for this purpose and to reduce the impact of large but erratic counts of some species – but the latter can now be catered for by dispersion weighting. Whilst this will eliminate the need for transformation in some cases, it will still be required in others (Clarke KR, Tweedley JR, Valesini FJ 2014, *J Mar Biol Ass UK* 94: 1-16), to down-weight large counts which are also consistent. (The example there is of counts of small-bodied fish species, and demonstrates the usefulness of shade plots – seen earlier in this section – in determining whether/what transform may be needed after dispersion weighting.)

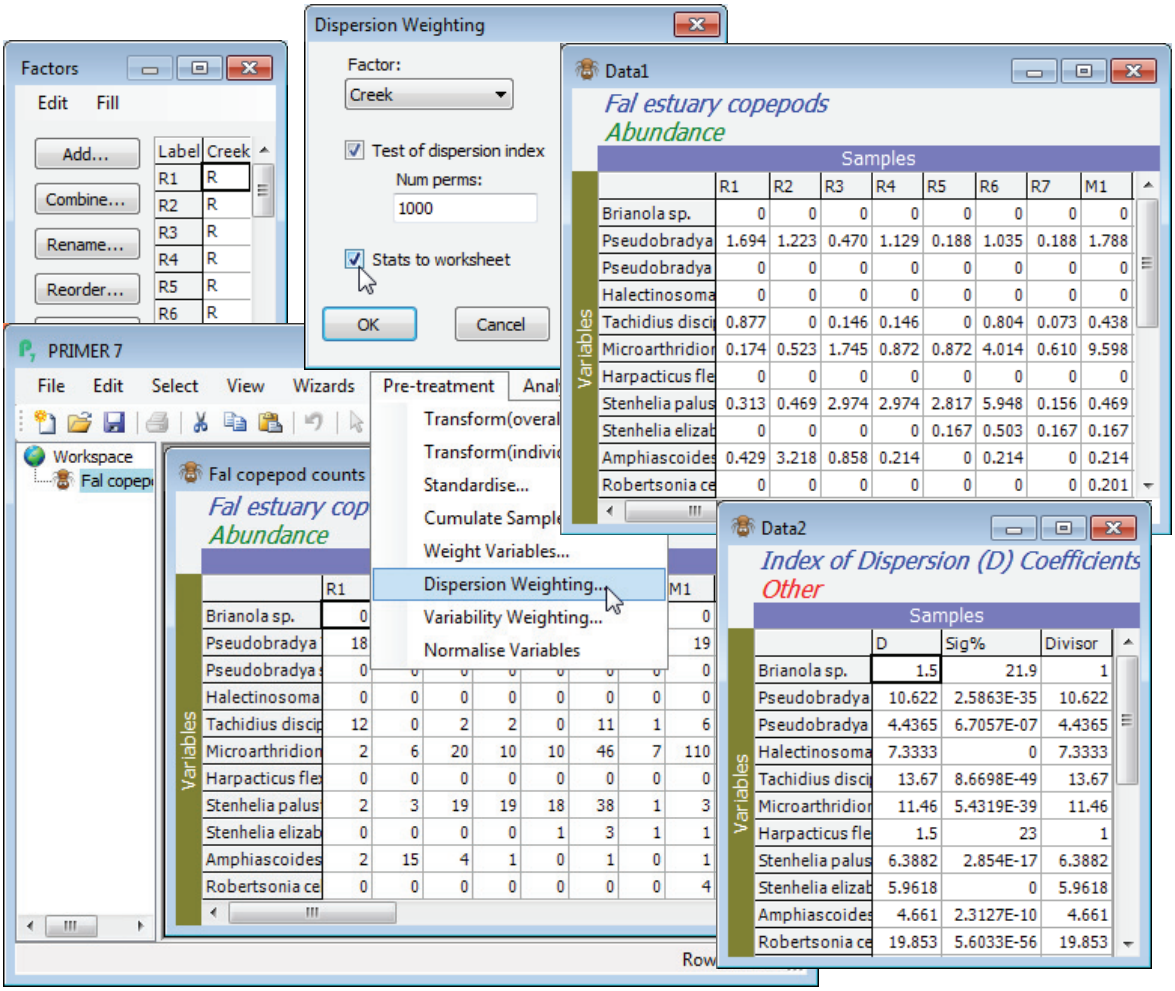
v7

Chapter 9 of CiMC also discusses generalising the dispersion weighting concept to data which are not strict counts, but are density, area cover or biomass, etc. For 'quantity' data of this type, on a common measurement scale, it can still make sense to apply dispersion weighting, e.g. colonial species in large patches can have high variability for their mean area, over replicate quadrats (as measured by grid intersections, perhaps), and thus less inherent reliability than individual small-bodied, motile species with the same mean area cover. However, a dispersion index of 1 no longer has meaning (values depend on the measurement units) and permutation testing of $\overline{D} = 1$ thus also makes no sense. The PRIMER 7 dialog for **Pre-treatment>Dispersion Weighting** now gives a tick box not to perform this test, and division of entries by $\overline{D}$ then takes place whatever its value.

(Fal estuary copepods)

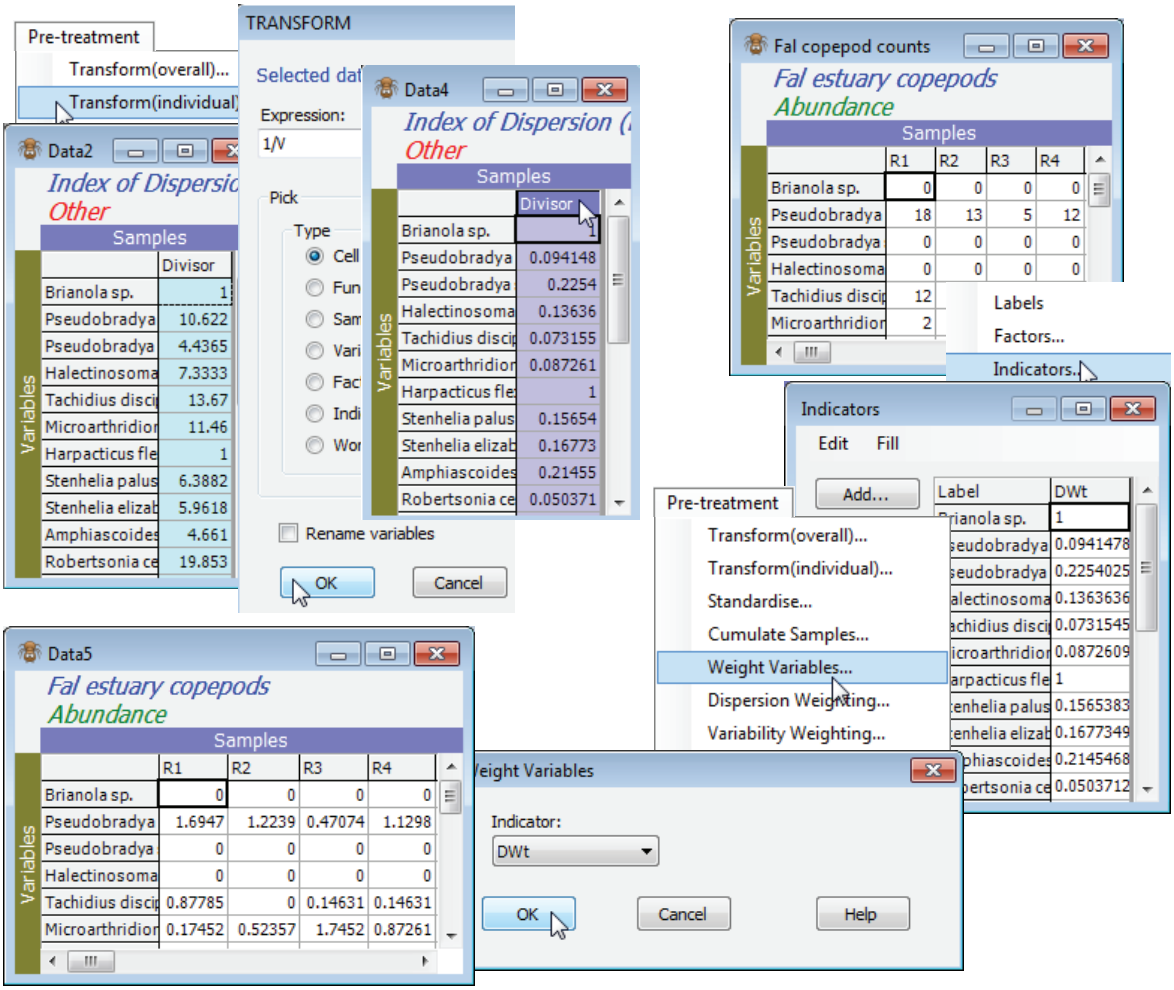
Sediment copepod assemblages (and other fauna) from five creeks of the Fal estuary, SW England, were analysed by Somerfield PJ, Gee JM, Warwick RM 1994, *Mar Ecol Prog Ser* 105: 79-88. The sediments of this estuary are characterised by high and varying concentrations of heavy metals, a result of tin and copper mining over hundreds of years. The copepod data consist of 23 species found in 27 samples, consisting of 5 replicate cores spanning each creek (Mylor: M1-M5; Pill: P1-P5; St Just: J1-J5; Percuil: E1-E5; and 7 from the largest creek, Restronguet: R1-R7). These are in directory C:\Examples v7\Fal benthic fauna, worksheet Fal copepod counts(.pri), with a factor *Creek* identifying samples from the 5 creeks. There are also environmental cores (of silt/clay ratios, heavy metals etc) matching these 27 sample locations, held in an Excel file Fal environment(.xls), plus nematode densities, macrofaunal counts and biomass, and associated aggregation files.

File>Open the copepod data and take **Pre-treatment>Dispersion weighting>**(Factor: Creek) & (✓Test of dispersion index) & (Num perms: 1000) & (✓Stats to worksheet). The Data1 sheet gives the dispersion weighted counts, which are either ready to go into the **Analyse>Resemblance** step of the next section, or could be mildly transformed before they do so, as shown earlier with **Pre-treatment>Transform(overall)>**(Transformation: Square root). There seems little need for the latter, however, since the dispersion weighting has already succeeded in downweighting the larger, erratic counts coming from *P. littoralis*, *R. celtica*, *E. gariene* and *T. discipes* and the somewhat less erratic *P. curticorne* and *M. falla* – the matrix Data1 now has no dispersion-weighted 'counts' in double figures, and the subsequent untransformed analysis will not be dominated by a small set of species. In three columns, Data2 gives: the mean dispersion indices $\overline{D}$ for each species; the evidence for clumping (i.e. the % significance level for a test of $\overline{D} = 1$); and the actual divisor used for that species row, which is 1 if the test does not reject this hypothesis at 5% (or better). Thus, *T. discipes* values are divided by 13.67 but *Brianola sp.* remains unchanged, though $\overline{D} = 1.5$. You might now like to run the routine again for the Fal nematode abundance file, which inspection shows must be numbers scaled up to a density, not real counts (e.g. there are no entries of 1!). The tick box for the test must be unchecked, the resulting $\overline{D}$ values are all >>1, but weighting by $\overline{D}$ is still justifiable.



Other variable weighting

There are other cases in which variables (species) might need prior weighting, e.g. when a species is known to be often misidentified, its contribution (and those of the species it is mistaken for) can be reduced by multiplying the entries in the two species through by some downweighting constant. This is achieved by placing weights for each species in an Indicator (see Section 2) and taking **Pre-treatment>Weight variables**, supplying the indicator name. In this context, most weights would be 1, with a value less than 1 used for downweighting less-reliably identified species (the default weight could be 100, or any number, since similarities such as Bray-Curtis are invariant to a scale change). A further context in which this routine might be useful is to convert counts to approximate biomass, using a known average weight of an individual of each species. Also dispersion weighting is seen just to be another case of variable weighting, with weights as the reciprocal of the Divisor column. You might like to demonstrate this for the Fal copepod counts example above, by selecting or highlighting the Divisor column from Data2 then take **Pre-treatment>Transform(individual)>** (Expression:1/V), highlighting the new column and copying (Ctrl-C) to the clipboard; opening Fal copepod counts, **Edit>Indicators>Add**>(Add indicator named:DWt), highlighting that blank new column and pasting (Ctrl-V); and finally **Pre-treatment>Weight Variables**>(Indicator:DWt). The resulting matrix should be identical to Data1. Save the workspace as Fal ws for later use.



Mixed data types

Another example might be in attempting to reconcile two different types of data in the same matrix, e.g. counts of motile organisms and area cover of colonial species. These cases can be problematic. One solution is to use a similarity measure such as the Gower coefficient, which scales the range of each species across samples to be identical, but this generally performs badly because very rare species are given the same weight as very common ones. A preferable alternative is to use Bray-Curtis similarity as usual, but prior to that **Weight Variables** to convert counts into approximate area cover, species by species, or both counts and area cover into a rough estimate of biomass, or even just to balance the two sets of variables against each other in some arbitrary way, e.g. give the cover numbers 10 times as much weight, or 10 times less weight, keeping the counts unchanged, and see what difference it makes to the analysis. (See also the discussion on p5-19 of CiMC.)

Variability weighting

v7

Pre-treatment>Variability Weighting is a new option in PRIMER 7, which bears similarities to the idea of Dispersion Weighting. This was introduced by Hallett CS, Valesini FJ, Clarke KR 2012, *Ecol Indicat* 19: 240-252 in a context where the variables were 'health indices' of fish communities and is exemplified here in a comparable case of a 'biomarker' suite, measured on individual fish, from locations with putatively differing contaminant impacts. Such indices typically behave more like environmental-type variables, with differing measurement scales and without the presence/absence structure of community matrices, so that transformation, normalisation and then a distance resemblance measure (e.g. Euclidean) would be appropriate (see earlier this section). The downside of normalising is, however, that all variables are essentially given equal weight in that calculation – but how else can one sum variable contributions over different units other than to shrink or stretch their scales to a common 'spread' (SD of 1)? (The location shift involved in normalising is actually irrelevant as far as distance measures such as Euclidean or Manhattan are concerned, because they are a function only of differences between sample values for each variable, so a subtracted constant disappears – the key thing is only the scale change to each variable). One possible answer is to scale each variable to a common spread (e.g. SD of 1) of its replicates within groups, not the full set of values across the groups (where the groups are the combinations of site, time etc). As with Dispersion Weighting, the idea is that some indices may be inherently less reliable than others, with erratic values for genuinely independent replicate observations within groups, so that it is desirable to give more weight to variables with lower (average) replicate variability. The variables now, though, are no longer 'quantities' – indeed after some transformations (e.g. log) they may take negative values, the mean may even be zero and dividing by the Index of Dispersion (ratio of variance to mean) will make no sense. Instead, the **Variability Weighting** dialog offers a range of possible rescalings of replicates, by: •Pooled SD (as would be calculated from 1-way ANOVA, by square-rooting the residual variance estimate, the logically best option for normally distributed variables with common replicate variance across groups); •Averaged SD (a simple mean of SDs computed separately for each group); •Averaged range (mean of the separate ranges – if used with Manhattan distance this is a more subtle version of the Gower coefficient, see Section 5); and •Averaged IQ range (mean of the inter-quartile ranges for each group, potentially a more relevant spread measure than SD for non-normally distributed – but continuous – replicate observations). As with Dispersion Weighting, all samples for each variable are simply divided through by their own averaged replicate 'spread', a new sheet formed and the divisors given in a Results window.

(Biomarkers for N Sea flounder)

The directory C:\Examples v7\N Sea biomarkers holds a data sheet N Sea flounder biomarkers(.pri) of biochemical and histological biomarkers measured concurrently on flounder caught at 5 North Sea sites (labelled S3, S5, S6, S7 and S9), running on a putative contaminant gradient from the mouth of the Elbe (S3) to the Dogger Bank (S9). It is a multivariate study because all variables were measured on each of 50 (pools of) fish, consisting of 10 sample pools from each site. This was part of a larger practical workshop on assessing 'biological effects' techniques for detecting pollution in the marine environment – the IOC Bremerhaven workshop (Stebbing ARD, Dethlefsen V, Carr M (eds) 1992, *Mar Ecol Prog Ser* 91, special issue). The 11 biochemical and sub-cellular variables measured: EROD, lysosomal acridine orange (LYS AO), lysosomal neutral-red retention (LYS NRR), pinocytosis (PINOCYT), oxyradicals (OXYRAD), endoplasmic reticulum (END RET), N-ras, ubiquitin, cathepsin D (CATH D), tubulin and lipid vacuoles (LIPID VAC), which are a mixture of continuous, heavily discretised and (one) presence/absence variables.

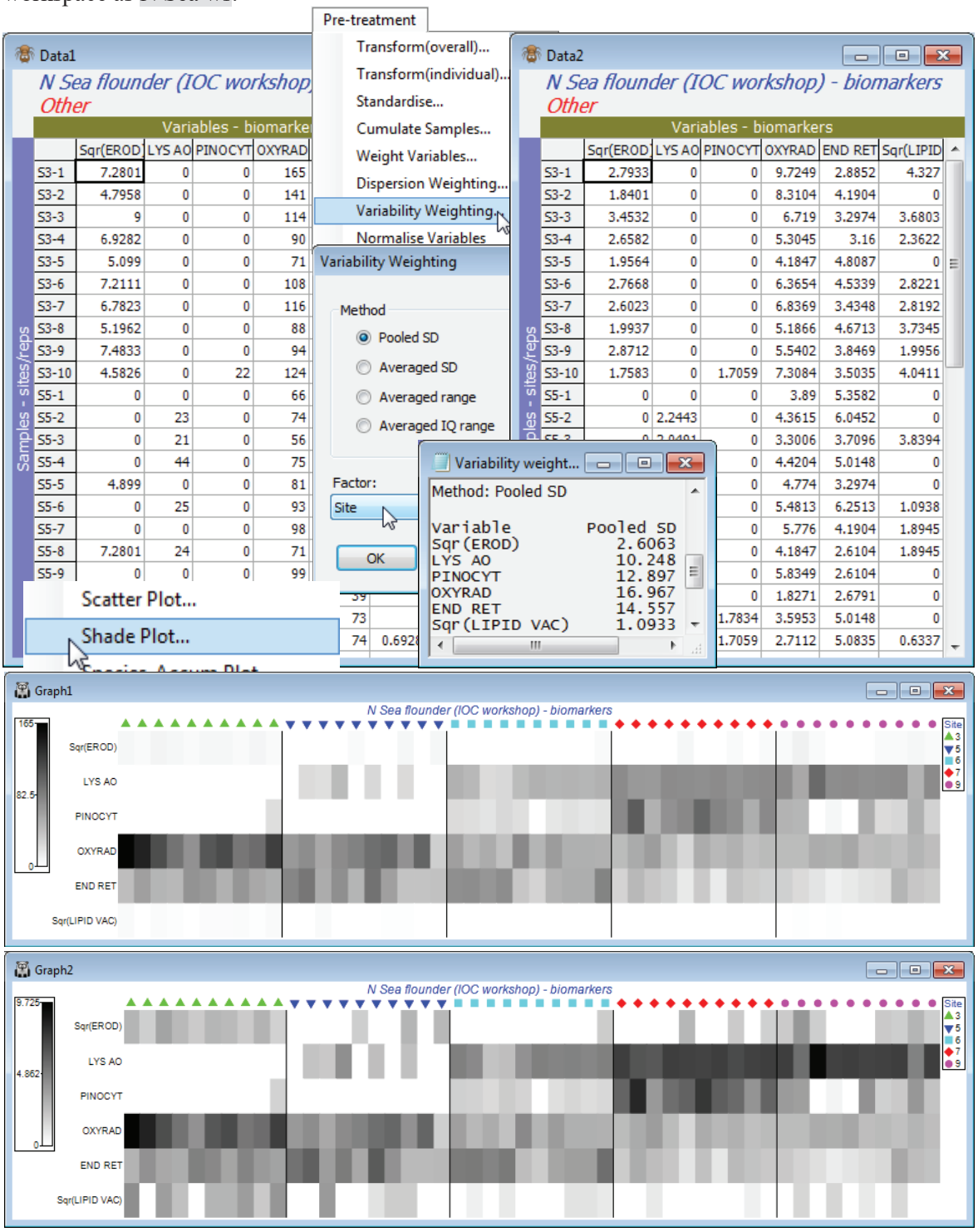
Open the N Sea flounder biomarkers sheet; note this time the 11 variables are columns and the 50 samples the rows, with a defined *Site* factor. Highlight, then select, only the 6 continuous variables EROD, LYS AO, PINOCYT, OXYRAD, END RET and LIPID VAC. They are all statistically well-behaved variables without strong outliers though the replicates for EROD and LIPID VAC are somewhat right-skewed, so it might be beneficial to highlight these two and take **Pre-treatment>Transform(Individual)**>(Expression:Sqr(V)) for a mild square root transform of those variables.

v7

Pre-treatment>Variability Weighting>(•Pooled SD) & (Factor:Site) then results in a sheet in which the columns (variables) are just divided by the pooled SD values given in the results window (Variability weighting1). The Data2 sheet is now ready to go into the **Analyse>Resemblance**> (Measure•Euclidean distance) routine of the next section. (Note that, though the Pooled SD divisor is really designed for continuous data, nothing goes dramatically wrong by leaving all 11 variables in the computations, though the discrete variables need to be at least ordered categories, as here).

v7 Finally, the effect on the relative contributions of each of the 6 variables to the Resemblance calculations, before and after the Variability Weighting, can be neatly seen by submitting both Data1 and Data2 to **Plots>Shade Plot**. The (transformed) EROD and Lipid variables would clearly be largely ignored without the rescaling but, less trivially, the variable clearly given greater weight (darker shading) by the rescaling is LYS AO, seen to have consistently high or consistently low values for replicates from the same site, in the original plot. The plots have been slightly modified, simply to accentuate the site differences on the *x* axis, using the extensive menu choices, e.g. for

v7 labels and adding symbols, **Graph>Sample Labels & Symbols**>(Symbols:(✓Plot) & (✓By factor) & Site) and taking off the tick box for Labels. This is a generic menu that covers many types of plot and will be seen again in Section 6 and beyond. Also, there are Special menus applying only to individual types of plot and one of those is demonstrated by the dividing lines between sites, which are produced by **Graph>Special>Reorder**>(Samples:Constraint•Factor groups:Site). The many other options on this Special menu for the shade plot routine are discussed in Section 10. Save the workspace as N Sea ws.



Cumulating samples

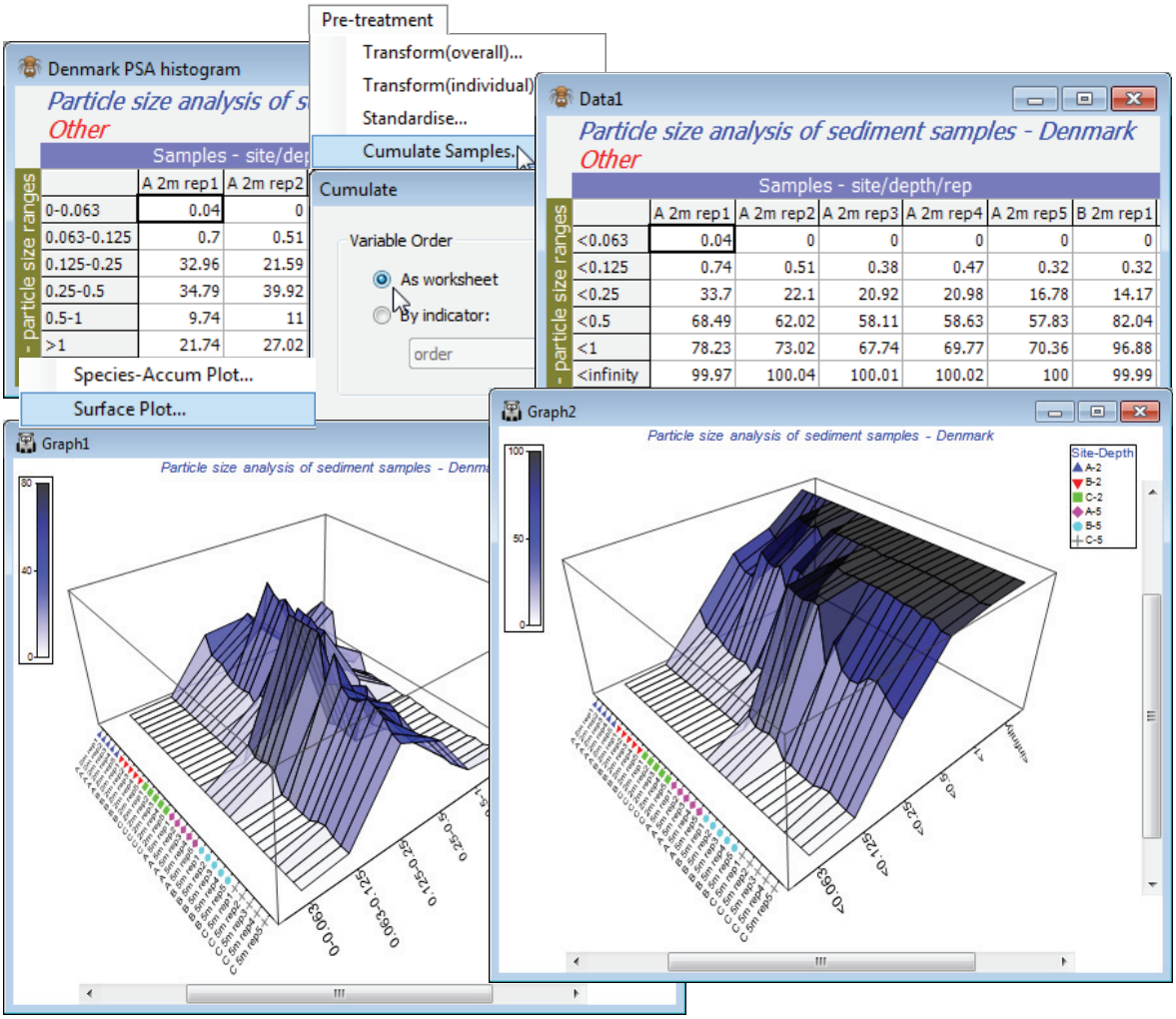
The remaining option on the Pre-treatment menu is **Cumulate samples**, which successively adds up the entries across variables, separately for each sample. It is only appropriate when all variables share a common measurement scale, and when the order in which they are listed is meaningful; it is thus not relevant standard species-by-samples data. It may be useful in analysing arrays in which variables are different body-size categories of a single species, or different particle sizes classes in Particle Size Analysis (PSA) etc, and entries are the frequencies or quantities of each size class in each sample (see CiMC, Chapter 8). Such data is typically analysed by univariate methods, fitting parametric particle-size distributions and comparing parameter estimates over samples. That can be problematic: histograms do not fit the models, summary statistics like mean and variance do not capture features such as bimodality, tests are incorrect because the data are not real frequencies, it is difficult to synthesise many such samples etc. This can be side-stepped by multivariate analysis, defining the similarity of pairs of size-class distributions. To take into account ordering of the sizes, when histograms are not smooth, it may sometimes be preferable to compare pairs of cumulative distributions (sample *distribution functions*) rather than the histograms (sample *density functions*).

(Particle sizes for Danish sediments)

Sediment particle size data from 6 size ranges at 3 sites (A, B, C), at 2 depths (2m and 5m), and 5 replicate samples from each site/depth combination are in C:\Examples v7\Denmark PSA\Denmark PSA histogram. The data are already standardised to % composition – if not, you would need **Pre-treatment>Standardise**>(Standardise•Samples)&(By•Total). So, take **Pre-treatment>Cumulate Samples>**(Variable Order•As worksheet). The variable labels are now inaccurate, so replace them by copying/pasting cumulative from **Edit>Indicators** and into the **Edit>Labels>Variables** dialog.

Surface plots

The smoothing effect can best be seen by **Plots>Surface Plot** of both Data1 and Data2 sheets – a further graph type new to PRIMER 7, but which should be reserved only for cases such as this, where there is genuine ordering of the variables. Symbols can be added to the sample axis with **Graph>Sample Labels & Symbols**, and colour shading introduced (as it could have been with the above Shade Plots) by **Graph>Special>Key**. Save the workspace as Denmark ws for use later.



5. *Resemblance*: similarities, dissimilarities and distances

Resemblance matrices

Fundamental to the operation of PRIMER and (explicitly or implicitly) any fully multivariate analysis, is an appropriate definition of resemblance between every pair of samples, based on whether the suite of recorded variables (species, environmental variables, biomarkers, particle-size classes or whatever) take similar or dissimilar values. What is meant by 'similar' is a function of the context and purpose of the analysis, and PRIMER 7 gives nearly 50 definitions to choose from (many are covered by the general reference work Legendre P & Legendre L 2012, *Numerical ecology, 3rd English ed*, Elsevier, called L&L from now on). Within PRIMER, similarity is taken to range over 0 to 100 (perfect similarity), dissimilarity is the complement (100 – similarity), whereas distance ranges from 0 to infinity. PRIMER 7 uses the term *Resemblance* to cover all three concepts: •Similarity, •Dissimilarity or •Distance, and also a number of specialised coefficient types which are useful to distinguish separately: •Distance[2]; •Correlation (which is defined over the range –1 to 1 and is therefore not directly a similarity, though it may be transformed into one in at least two different ways – see the Transform option in Section 11); •R (the pairwise ANOSIM R statistic – see Section 9); and •Rank (where similarities or dissimilarities are turned into ranks, i.e. the positive integers, with averaged values for any tied ranks – which can be used directly as a distance matrix. The unifying structure here is that these are all pairwise coefficients and they are all symmetric (the resemblance of samples 1 and 2 is the same as that of 2 and 1), so resemblances between every pair of samples form a lower triangular matrix, with no diagonal. They are displayed with the upper triangle absent and the specific Type as the second heading in the sheet window, so it should always be clear when the active window is a resemblance matrix and when it is a data sheet. (This matters because the available menu options change with the active window type).

v7

Standard resemblance choices

A detailed discussion of the competing properties of different resemblance matrices is outside this manual's scope (see L&L, CiMC Chapters 2 & 16, or Clarke KR, Somerfield PJ, Chapman MG 2006, *J Exp Mar Biol Ecol* 330: 55-80). Novice users are recommended to take one of the main options (the defaults): Bray-Curtis similarity for biological assemblage data; Euclidean distance (having first normalised) for physico-chemical, biomarker or morphometric data etc, in which variables are not on comparable ranges or the same measurement scale at all; and (non-normalised) Euclidean distance for body- and particle-size histograms (first standardised), growth curves etc.

Bray-Curtis similarity

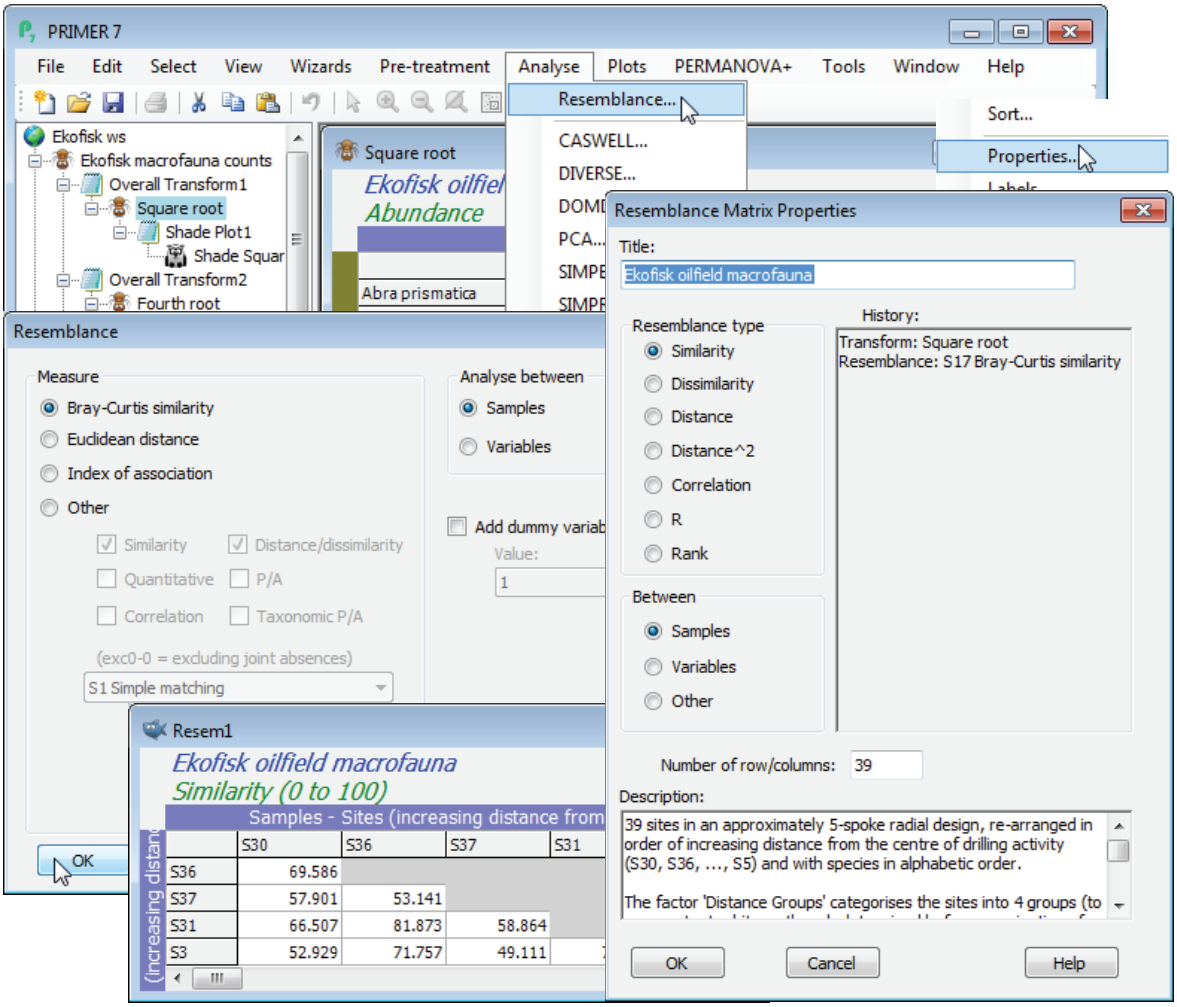
The most commonly-used similarity coefficient for biological community analysis, because it obeys many of the 'natural' biological guidelines in a way that most other coefficients do not (see CiMC), is the Bray-Curtis similarity, defined between samples 1 and 2 as:

$$S_{17} = 100\left(1 - \frac{\sum_i |y_{i1} - y_{i2}|}{\sum_i y_{i1} + \sum_i y_{i2}}\right) \equiv 100.\frac{\sum_i \min\{y_{i1}, y_{i2}\}}{(\sum_i y_{i1} + \sum_i y_{i2})/2}.$$

The two forms may not look identical but they are! Here y_{i1} is the count (or biomass, % cover, …) for the ith (of p) species from sample 1, and $\Sigma_i(\ldots)$ denotes summation over those species. Original references to coefficient definitions are not given here (nomenclature is always a source of debate!) – see L&L, whose numbering scheme is followed where possible, hence S_{17} for Bray-Curtis.

Open the workspace C:\Examples v7\Ekofisk macrofauna\Ekofisk ws from earlier, and click on the Square root counts sheet (obtained earlier with **Pre-treatment>Transform(overall)>**Square root). Take **Analyse>Resemblance>**(Measure•Bray-Curtis similarity) & (Analyse between•Samples), which are the defaults for this data type. A lower triangular matrix is produced, Resem1, which you should rename B-C on sq rt. **Edit>Properties** (or right-clicking over the matrix to get **Properties**) shows it is of Resemblance type•Similarity from 39 samples. The History box carries through the knowledge of how it was created to a subsequent Cluster or MDS ordination plot. This box is not user-editable, though the Title and Description boxes can be altered; changes to the Title are carried forward to a subsequent plot but not backward to the data sheet Square root.

Now repeat **Resemblance** directly on the original Ekofisk macrofauna counts, without the **Pre-Treatment** transformation. PRIMER tries to help – a warning message appears that no transform has been applied; community matrices usually require some transformation before calculating Bray -Curtis (though you can happily ignore this warning if you are interested in the pattern of the few most dominant species only). **Cancel** the calculation and resave the Ekofisk ws workspace.



Zero-adjusted Bray-Curtis

A simple modification to the Bray-Curtis coefficient adjusts its behaviour as samples become vanishingly sparse. Standard Bray-Curtis is undefined for two samples containing no species, and can fluctuate wildly for near-blank samples – two samples containing just a single individual can fluctuate between 100% similarity if the individuals are from the same species, to 0% if they are not. The zero-adjusted Bray-Curtis coefficient (Clarke KR, Somerfield PJ, Chapman MG 2006, *J Exp Mar Biol Ecol* 330:55-80; also CiMC, Chapter 16) damps down this behaviour – analogously to the addition of the constant 1 in the log(1+x) transformation (to cater for x=0) – by adding +2 to the denominator of the ratio in S_{17}. A simple way of viewing this is as adding a 'dummy species' to the matrix, taking the value 1 for all samples. This forces two samples with no content to be 100% similar (they share the dummy species) and two samples with a single real individual now have some similarity, whether that species is shared (100%) or not (50%). It is clear that once there are a modest number of individuals, in either sample, then the adjustment makes no difference. It can only come into force when the assemblage is virtually denuded, and should only be applied if it makes biological sense to regard two blank samples as 100% similar, because both are denuded *from the same environmental cause*. If blank samples can be present in very different treatments/ sites etc, because of small sample sizes and highly clustered spatial distributions of organisms, it is unwise to use the zero-adjustment – instead, remove the blanks and use standard Bray-Curtis.

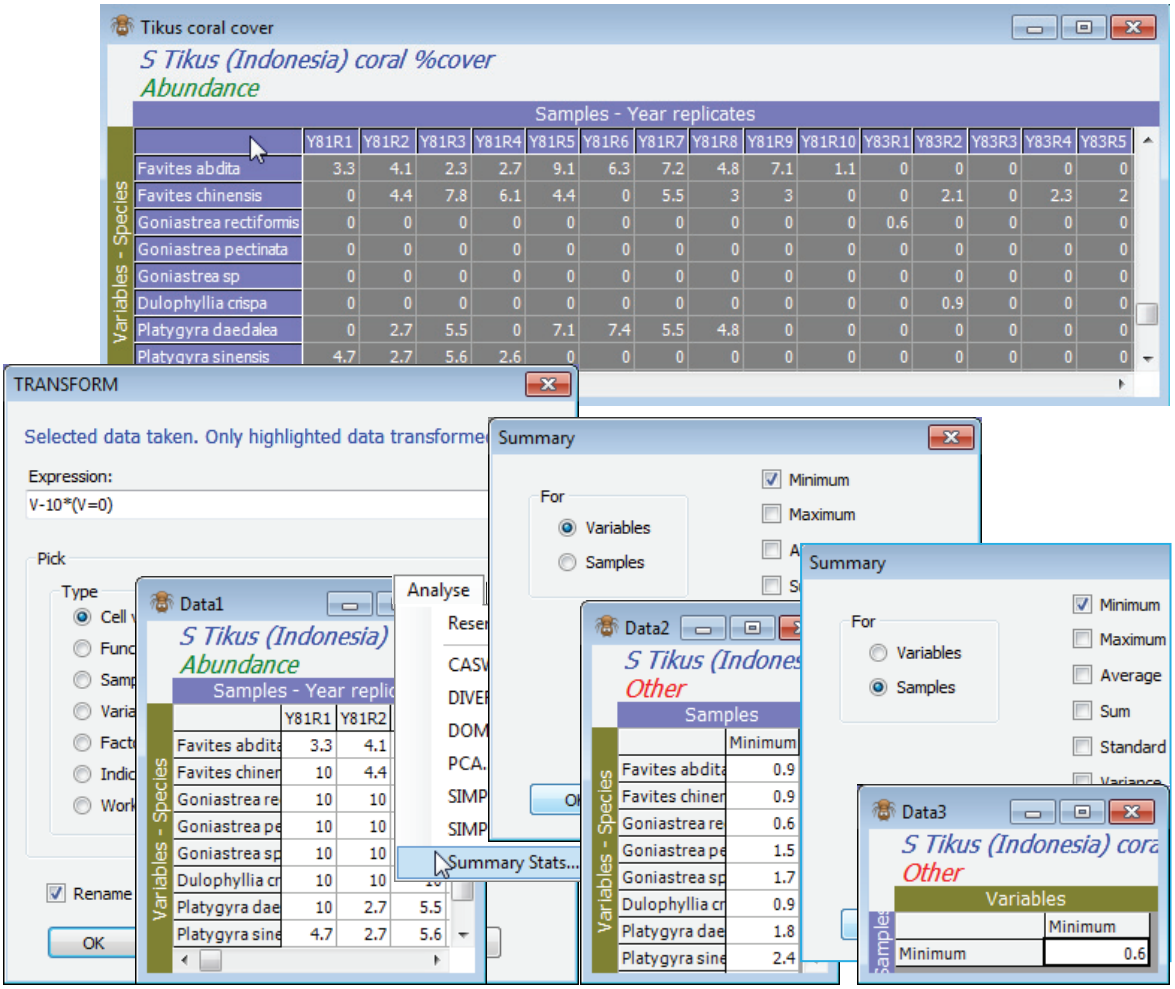
The adjustment is made by taking: (✓Add dummy variable)>(Value:1) in the Resemblance dialog. The constant 1 is appropriate to integer counts, being the lowest non-zero value attainable. This is true whether the data sheet has previously been transformed or not (the constant remains 1 under any power transform). For data on biomass, % area cover etc, the value could sensibly be chosen similarly as the lowest non-zero entry likely to be recorded (again the analogy with the log(c+x) transform is appropriate). 'Adding a dummy variable' can be carried out with other resemblance measures, but will only be effective for those coefficients which, like Bray-Curtis, treat joint absences of species as uninformative (e.g. Kulczynski, Czekanowski mean character difference, Canberra etc). It is not given as an option for data type Environmental (it makes no sense then).

(Tikus Island coral cover)

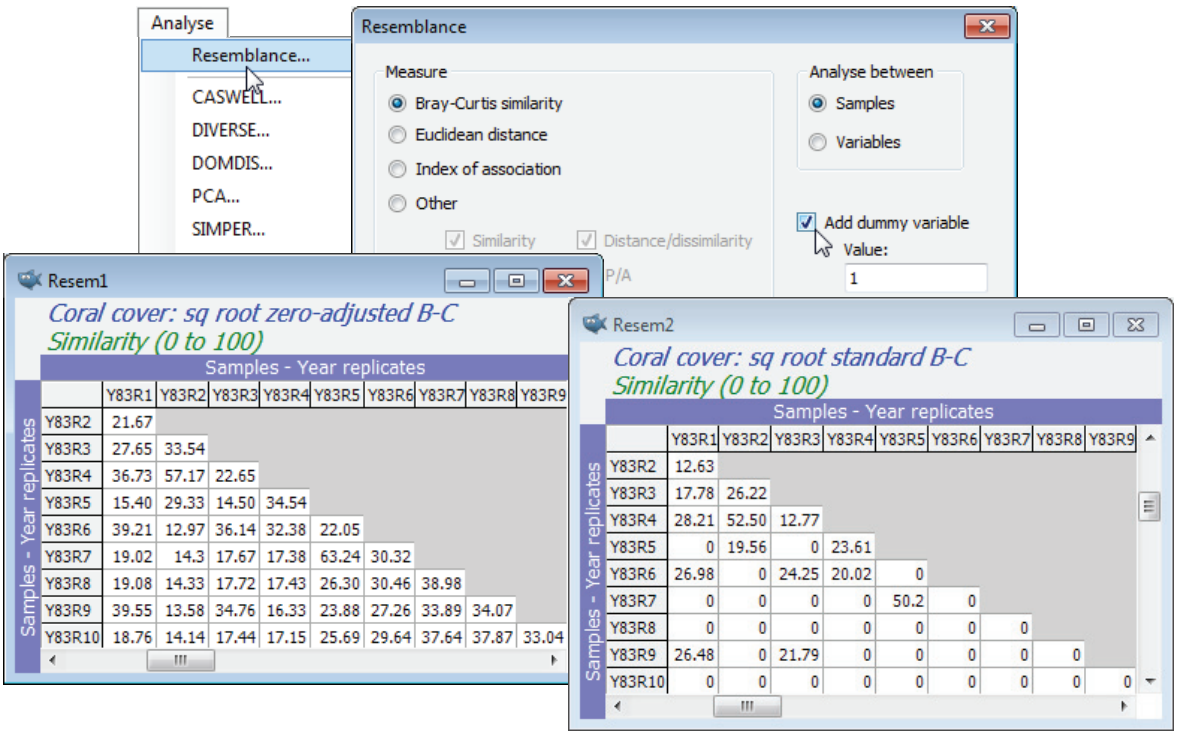
Data on coral communities at a site in Tikus Island, Thousand Islands, Indonesia, over the years 1981, 83, 84, 85, 87 and 88, were reported by Warwick RM, Clarke KR, Suharsono 1990, *Coral Reefs* 8: 171-179. Ten replicate transects were examined each year, and the data is the length of intersection of a transect (as a percentage of transect total length) by each of the 58 coral species identified, file Tikus coral cover in directory C:\Examples v7\Tikus corals. The region was subject to a coral bleaching event in 1982 (probably El Niño related), so that the 1983 samples are very denuded of live coral – this is a classic situation in which a zero-adjusted Bray-Curtis similarity is likely to be useful, and this example is discussed in detail in the Clarke *et al* 2006 paper mentioned above. A dummy value of 1 is a natural choice here because the smallest non-zero entries for each species are about 1%, or marginally less. To see these entries, highlight the whole array, take say **Pre-treatment>Transform(individual)**>(Expression:V-10*(V=0)) and enter the resulting sheet to

v7 **Analyse>Summary Stats**>(For•Variables)&(✓Minimum). This works because the BASIC syntax expression computed on the value V in every cell, V-10*(V=0), returns either -1 (true) or 0 (false) for V=0, multiplies this up to -10 or 0, so when subtracted from V returns +10 in any cell which is

v7 zero and leaves non-zero values alone. **Summary Stats** then finds the minimum for each species. (10 in the expression could be replaced by any large number). If you run the Summary Stats again, this time (For•Samples)&(✓Minimum) you will get the lowest non-zero entry in the whole matrix.



v7 As in the previous section, **Plots>Shade Plot** readily shows that a (mild) square root transform is necessary to avoid the resemblance calculation being dominated by just a couple of species with occasionally very large %cover values. So, after **Pre-treatment>Transform(overall)>**Square root, take **Analyse>Resemblance>**(Measure•Bray-Curtis similarity) & (Analyse between•Samples) & (✓Add dummy variable>Value: 1), this dummy value of 1 being equally suitable after any power transformation or reduction to presence/absence (1/0). By repeating this calculation on the square-rooted data, but without the dummy variable, a quick glance at the two resemblance matrices shows the dramatic effect of the zero-adjustment here, e.g. among the 1983 replicates. (This translates into substantial differences in the clustering, MDS ordination, ANOSIM tests etc, see Fig. 16.7, CiMC). **File>Save Workspace As>**(File name:Tikus ws) in the C:\Examples v7\Tikus corals directory.



Resem1

	Y83R1	Y83R2	Y83R3	Y83R4	Y83R5	Y83R6	Y83R7	Y83R8	Y83R9
Y83R2	21.67								
Y83R3	27.65	33.54							
Y83R4	36.73	57.17	22.65						
Y83R5	15.40	29.33	14.50	34.54					
Y83R6	39.21	12.97	36.14	32.38	22.05				
Y83R7	19.02	14.3	17.67	17.38	63.24	30.32			
Y83R8	19.08	14.33	17.72	17.43	26.30	30.46	38.98		
Y83R9	39.55	13.58	34.76	16.33	23.88	27.26	33.89	34.07	
Y83R10	18.76	14.14	17.44	17.15	25.69	29.64	37.64	37.87	33.04

Resem2

	Y83R1	Y83R2	Y83R3	Y83R4	Y83R5	Y83R6	Y83R7	Y83R8	Y83R9
Y83R2	12.63								
Y83R3	17.78	26.22							
Y83R4	28.21	52.50	12.77						
Y83R5	0	19.56	0	23.61					
Y83R6	26.98	0	24.25	20.02	0				
Y83R7	0	0	0	0	50.2	0			
Y83R8	0	0	0	0	0	0	0		
Y83R9	26.48	0	21.79	0	0	0	0	0	
Y83R10	0	0	0	0	0	0	0	0	0

Euclidean distances

Euclidean distance, an appropriate measure for environmental (and other) data types, is defined as:

$$D_1 = \sqrt{\sum_i (y_{i1} - y_{i2})^2}$$

where the y_{i1} & y_{i2} result from pre-treatment by transformation (sometimes) and subsequent normalisation (often). The outcome is a triangular distance matrix, which orders in the opposite direction to similarity: high similarity = low distance (= low dissimilarity). Note, however, that the user does not have to worry about which way round the resemblances are ordered: all routines will utilise the information given in the Resemblance type to make sensible choices.

Re-open the Ekofisk workspace Ekofisk ws from the \Ekofisk macrofauna directory; you should have available the transformed and normalised environmental data (Data4 perhaps) from Section 4, on which to calculate Euclidean distance. The **Analyse>Resemblance** dialog box now gives the default as Measure•Euclidean because Data type has been defined as Environmental, so you can take the defaults here. The result is a resemblance matrix of type Distance; the History box on the **Edit>Properties** dialog shows its derivation as Euclidean distance on normalised data. Compute Manhattan distance also (see next page) and rename the sheets as Euclid and Manhattan by clicking (twice, slowly) on their default *Resem* names in the Explorer tree. Most other measures in the lists below are not suitable for normalised environmental data but are designed for positive 'quantities'.

Index of Association

The remaining of the three choices in the initial list, Measure•Index of Association, is essentially Whittaker's index of association, which when calculated on samples (the default is always Analyse between•Samples) is simply just Bray-Curtis similarity on a sample-standardised matrix. [You might like to check this with the following sequences on the original Ekofisk macrofauna counts :
(i) **Pre-treatment>Standardise**>(Standardise•Samples)&(By•Total) then **Analyse>Resemblance** >(Measure•Bray-Curtis similarity)&(Analyse between•Samples), compared with
(ii) **Analyse>Resemblance**>(Measure•Index of Association)&(Analyse between•Samples).]

v7

The Index of Association is not therefore in this main list for its use on sample similarities but because it is the primary means of computing similarities among species, in their behaviour over the full set of samples. Importantly therefore, (Measure•Index of Association) almost always needs to be used with (Analyse between•Species), and the measure is then defined, over (0,100), as:

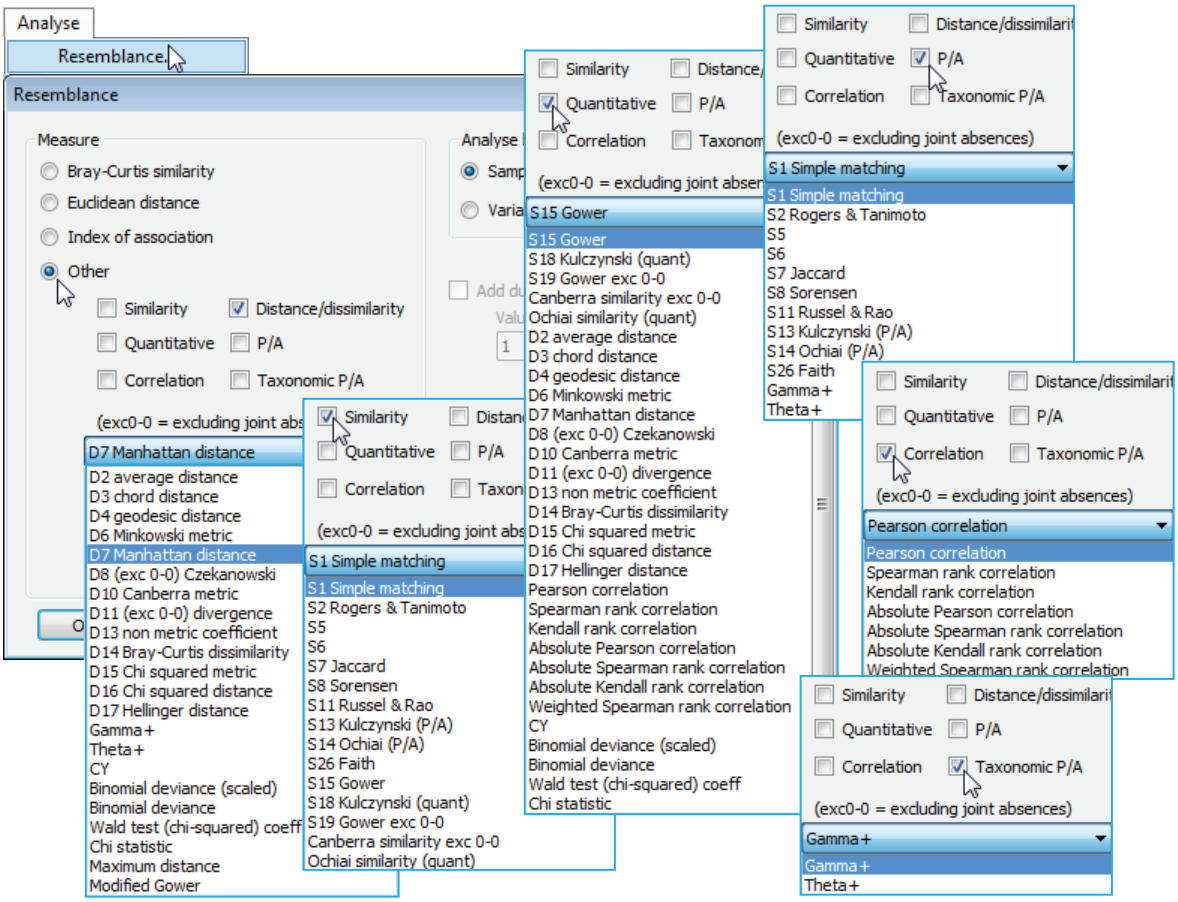
$$IA = 100\left[1 - \frac{1}{2}\sum_j \left| \frac{y_{1j}}{\sum_j y_{1j}} - \frac{y_{2j}}{\sum_j y_{2j}} \right| \right]$$

v7 !

with 0 implying full 'negative' and 100 full 'positive' association of the two species (1 & 2 in the above equation). For its application as part of the new *coherent curves* method in PRIMER 7 see Section 10. Note that in PRIMER 6, the Whittaker coefficient was present only in its dissimilarity form (the D_9 of L&L) which is really a coefficient of dis-association since it takes larger values for samples with more differing communities. The previous nomenclature was therefore confusing and the index of association is now available in PRIMER 7 only as a similarity. Note also that all the definitions in the remainder of this section (up to the Analysing between variables box heading) are given in terms of resemblances among samples, the primary use for resemblance matrices.

Accessing other resemblance measures

PRIMER 7 allows the user choice of 44 other resemblances, firstly divided into two (mutually exclusive) types: Similarities (including L&L's S numbers) or Dissimilarities/Distances (including L&L's D numbers); then most of the same coefficients split instead into Quantitative or Presence/Absence measures; and finally two specialised groups of Correlation coefficients and Taxonomic-based P/A measures (the latter using an aggregation file of the type met in Section 2, on species relatedness). These are all accessed through the Measure•Other button and the drop-down list.



Distance measures

The distance measures defined by L&L and calculated by PRIMER 7 (in addition to D_1) are:

$$D_2 = \sqrt{\frac{1}{p}\sum_i (y_{i1} - y_{i2})^2}$$ average distance,

where the number of species p is fixed for all pairs of samples, so this is a constant multiple of Euclidean distance D_1 and will therefore give identical dendrograms, ordinations etc. (complete data is assumed for all these formulae, i.e. without missing values, though automatic adjustment to formulae under *pairwise elimination* of missing values is carried out for all measures, see later);

v7 !

$$D_3 = \sqrt{2\left(1 - \frac{\sum_i y_{i1} y_{i2}}{\sqrt{\sum_i y_{i1}^2 \sum_i y_{i2}^2}}\right)}$$ Orloci's chord distance;

$$D_4 = \arccos\left(1 - \frac{1}{2} D_3{}^2\right)$$ geodesic metric;

$$D_6 = \left(\sum_i |y_{i1} - y_{i2}|^r\right)^{1/r}$$ Minkowski metric,

where r can be specified by the user (note r=1 gives Manhattan, and r=2 Euclidean distance);

$$D_7 = \sum_i |y_{i1} - y_{i2}|$$ Manhattan distance,

whose use of absolute rather than squared differences confers slightly better robustness to outliers;

$$D_8 = \frac{1}{p_{12}} \sum_i |y_{i1} - y_{i2}|$$ Czekanowski's mean character difference,

in the form where p_{12} is the number of species that are not jointly absent in samples 1 and 2 (the changing denominator across pairs of samples, from excluding joint absences, can make a big difference to a coefficient's behaviour, so is indicated clearly by 'exc0-0' in the drop-down box).

$$D_{10} = \sum_i \frac{|y_{i1} - y_{i2}|}{(y_{i1} + y_{i2})}$$ Canberra metric of Lance & Williams,

which must exclude joint absences so that it can be defined, but is less useful than its averaged form, divided by p_{12}, found as Canberra similarity in the quantitative similarity list;

$$D_{11} = \sqrt{\frac{1}{p_{12}} \sum_i \left(\frac{y_{i1} - y_{i2}}{y_{i1} + y_{i2}}\right)^2}$$ Clark's coefficient of divergence,

also in the form in which double zeros are excluded from the summation and the divisor p_{12};

$$D_{15} = \sqrt{\sum_i \frac{1}{y_{i+}} \left(\frac{y_{i1}}{\sum_i y_{i1}} - \frac{y_{i2}}{\sum_i y_{i2}}\right)^2}$$ χ^2 (chi-squared) metric,

where $y_{i+} = \sum_j y_{ij}$, the sum across all samples of the entries for the ith species, and effectively the same, to within a constant, as the following;

$$D_{16} = \sqrt{\sum_i \frac{1}{y_{i+} / \sum_i y_{i+}} \left(\frac{y_{i1}}{\sum_i y_{i1}} - \frac{y_{i2}}{\sum_i y_{i2}}\right)^2}$$ χ^2 distance,

the implicit distance underlying Correspondence Analysis, which is seen to be a type of Euclidean distance, from samples which are standardised by their totals across species, and then inversely weighted by species totals across samples (the double standardisation being responsible for the practical difficulties χ^2 distance can have with rare species, for which the divisor is near zero); and

$$D_{17} = \sqrt{\sum_i \left[\sqrt{\frac{y_{i1}}{\sum_i y_{i1}}} - \sqrt{\frac{y_{i2}}{\sum_i y_{i2}}}\right]^2}$$ Hellinger distance, advocated by Rao,

the only omission above being D_{13}, which is simply the complement of Sørensen similarity, S_8.

'Modified Gower'

Anderson MJ, Ellingsen KE, McArdle BH 2006, *Ecol Lett* 9: 683-693 used Czekanowski's mean character difference (above) as their preferred distance measure after a specific transformation of the original counts, advocated for its interpretable properties, namely: $y' = \log(y) + 1$, unless $y = 0$, when $y' = 0$. Choice of the base for the logarithm explicitly scales how much weight the counts get in relation to the presence/absence structure. For example, base 2 gives the step from 0 (absence) to 1 (individual) the same weight as the step from 1 to 2, or from 2 to 4, or 4 to 8 etc. Base 10 gives 0 to 1 the same weight as 1 to 10, or 10 to 100 etc. Thus high bases give more weight to the presence/

absence structure. Thus, this work mainly concerns an added transformation choice rather than a new resemblance measure, but it is convenient to bundle the transformation with Czekanowski's measure into a single coefficient, which the authors called *modified Gower* (though note that it avoids one of the defining, and usually problematic, features of the Gower coefficient S_{19}, below – that of standardising each species by its range of values across the samples). It is important to stress that the transform applies only to genuine counts (without other initial standardising/transforming). For densities, biomass, cover etc, the logic breaks down: y values can be less than 1, for which the transformed y' can be <0. Thus high densities give positive values for y' but low densities can give negative y' and an even lower density (absence) will give $y' = 0$ – the transform is not monotonic! To avoid this, any y values in (0,1) are initially rounded down or up to 0 or 1 before computation but this changes the number of perceived absences. Unless you are clear about the implications, the safest course is to use Modified Gower only for real counts – for which it is designed!

Similarity to dissimilarity

L&L also assign D_{14} to Bray-Curtis dissimilarity, the complement of S_{17}, defined earlier. This is also provided in the Dissimilarity list since it is (very occasionally) useful to specify a dissimilarity rather than its complementary similarity – though normally PRIMER will take either form into any of its routines and interchange similarity and dissimilarity where it needs to. This interchange can be performed explicitly, though, if you wish (perhaps for outputting a matrix of one or other type), by taking **Tools>Dissim** which uses the relation $D + S = 100$ to convert from S to D or D to S.

Quantitative similarity measures

In addition to Bray-Curtis S_{17}, and its zero-adjusted modification, PRIMER 7 also calculates:

$$S_{15} = 100.\frac{1}{p}\sum_i \left[1 - \frac{|y_{i1} - y_{i2}|}{R_i}\right], \text{ where } R_i = \max_j\{y_{ij}\} - \min_j\{y_{ij}\} \qquad \text{Gower's coefficient,}$$

where standardisation is by the range R_i of values for the ith species over all samples (effectively by the maximum since the minimum will usually be zero), and thus shares with χ^2 distance the (generally undesirable) property that adding further samples can change existing similarities;

$$S_{18} = 100.\frac{\sum_i \min\{y_{i1}, y_{i2}\}}{2/[(1/\sum_i y_{i1}) + (1/\sum_i y_{i2})]} \qquad \text{Kulczynski similarity,}$$

which can be seen from the second form of S_{17} to be related to Bray-Curtis, replacing the arithmetic mean of the sample totals in the denominator of S_{17} with a harmonic mean;

$$S_{19} = 100.\frac{1}{p_{12}}\sum_i \left[1 - \frac{|y_{i1} - y_{i2}|}{R_i}\right] \qquad \text{Gower (excluding double zeros),}$$

which is S_{15} with the fixed total number of species in the matrix (p) being replaced by p_{12}, the number of non-jointly absent species in the two samples being compared – an important difference;

$$S^{Can} = 100.\left(1 - \frac{1}{p_{12}}\sum_i \frac{|y_{i1} - y_{i2}|}{(y_{i1} + y_{i2})}\right) \qquad \text{Canberra similarity,}$$

in the form used by Stephenson W, Williams WT, Cook SD 1972, *Ecol Monogr* 42: 387-415*,* not numbered by L&L but of more use for species data than its distance form (Canberra metric) D_{10}, because of the division by the variable species numbers p_{12} (i.e. excluding double zeroes);

$$S^{M-H} = 100.\left(1 - D_1'^2 \Big/ \left(\sum_i y_{i1}'^2 + \sum_i y_{i2}'^2\right)\right) \qquad \text{Morisita-Horn similarity}$$

where $'$ denotes that y's are sample-standardised before D_1 and the denominator are calculated; and

$$S^{Och} = 100.\frac{\sum_i \min\{y_{i1}, y_{i2}\}}{\sqrt{\sum_i y_{i1} \sum_i y_{i2}}} \qquad \text{quantitative Ochiai similarity,}$$

not defined by Ochiai as such, but it reduces to Ochiai's coefficient (S_{14}) when applied to P/A data. Clarke *et al* 2006 (see above for reference) construct this coefficient – which is an intermediate form between Bray-Curtis and Kulczynski, because it replaces the denominator with a geometric rather than arithmetic or harmonic mean – to illustrate that measures with reasonable properties are not difficult to invent, explaining the plethora of coefficients available in the literature!

Presence/ Absence similarities

There are numerous similarity measures defined for simple species lists, i.e. when the data consist only of presence (1) or absence (0) of each species in each sample. Any similarity defined between samples 1 and 2 must then be a combination of only four numbers: *a*, the number of species present in both samples; *b*, the number present in 1 but absent from 2; *c*, the number absent in 1 but present in 2; *d,* the number absent from both. Clearly, the coefficient must be symmetric in *b* and *c*, and the more biologically useful coefficients are also not a function of joint absences, *d*. There still remain a large number of options, of which PRIMER 7 calculates the following:

$S_1 = 100 \cdot \frac{a+d}{a+b+c+d}$ simple matching; $S_2 = 100 \cdot \frac{a+d}{a+2b+2c+d}$ Rogers & Tanimoto;

$S_5 = 25 \cdot \left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d} \right]$; $S_6 = 100 \cdot \frac{a}{\sqrt{(a+b)(a+c)}} \cdot \frac{d}{\sqrt{(b+d)(c+d)}}$;

$S_7 = 100 \cdot \frac{a}{a+b+c}$ Jaccard; $S_8 = 100 \cdot \frac{2a}{2a+b+c}$ Sørensen;

$S_{11} = 100 \cdot \frac{a}{a+b+c+d}$ Russell & Rao; $S_{13} = 50 \cdot \left[\frac{a}{a+b} + \frac{a}{a+c} \right]$ Kulczynski (P/A);

$S_{14} = 100 \cdot \frac{a}{\sqrt{(a+b)(a+c)}}$ Ochiai (P/A); $S_{26} = 100 \cdot \frac{a+(d/2)}{a+b+c+d}$ Faith.

A quantitative matrix input to one of these calculations will automatically be reduced to a simple array of 1's and 0's before computation. The most frequently met of the presence/absence measures are Sørensen, which is Bray-Curtis calculated on P/A data, and Jaccard – the definition shows how alike they are. In fact they are monotonically related (as one increases, so does the other), so the procedures in PRIMER which are based only on rank values of the coefficients (i.e. most of them: nMDS, ANOSIM, BEST, RELATE etc, in our largely non-parametric approach to resemblance matrix analysis) will give exactly the same outcome for these two coefficients.

Quantitative measures on P/A data

It is instructive to draw the other links between quantitative coefficients and the presence/absence measures they reduce to, when calculating them on a P/A matrix. Pure distance measures such as D_1, D_6, D_7 and D_{10}, which are not averaged in some way over the number of species, clearly cannot reduce to the dimensionless ratios in the P/A similarity definitions above. Similarly, D_{15}, D_{16}, S_{15} and S_{19} are not of interest in this context because they are not just functions of *a, b, c, d* for the two samples but bring in species for all other samples, in their species standardisations. However, the other quantitative measures mainly reduce to simple monotonic functions of four P/A similarities: S_1 (simple matching), S_7 (Jaccard), S_8 (Sørensen) and S_{14} (Ochiai P/A). Of course, as defined, the relationships will be between *D* and (1 – S/100). To be precise: D_2 reduces to the square root of the complement of $S_1/100$; both D_3 and D_{17} go to the square root of $2(1 - S_{14}/100)$, D_4 to $\cos^{-1}(S_{14}/100)$ and S^{Och} to S_{14}; D_8 reduces to the complement of S_7, D_{11} to the square root of that complement, and S^{Can} to S_7. As noted earlier, S_{17} reduces to S_8 and, finally, S_{18} goes to S_{13}.

In less technical description: average Euclidean distance (squared) is the natural counterpart of simple matching (they are both functions of the number of joint absences); chord, geodesic and Hellinger distance, and naturally quantitative Ochiai, all have an affinity to the P/A form of Ochiai; Czekanowski's mean character difference, the divergence coefficient and Canberra similarity all relate to Jaccard; Bray-Curtis reduces to Sørensen and, unsurprisingly, the quantitative and P/A forms of the Kulczynski coefficient converge, e.g. as strong transforms force the data towards P/A.

Demonstrate one of these points for the Ekofisk abundance data in the Ekofisk ws – which should still be open – by calculating Hellinger distance (D_{17}) on the presence/absence data produced from the macrofauna sheet, and comparing this with the Ochiai P/A coefficient (S_{14}). Thus:

a) With Ekofisk macrofauna counts as the active window, **Pre-treatment>Transform(overall)>** (Transformation: Presence/absence) to produce the P/A matrix, then renamed P-A (forward slash is not a permitted symbol in the Explorer tree, since these may sometimes be filenames);

b) On P-A, **Analyse>Resemblance>**(Measure•Other: D17 Hellinger distance) & (Analyse between •Samples), renaming the *Resem* sheet to Hell on P-A. [Do not take 'Add dummy variable' here – or routinely (always think carefully about it first!). It will have negligible effect here on relative distances because there are no denuded samples at all. However, the option is permitted with all measures and could make sense, in the presence of blank or near-blank samples (which are then required to have zero or near-zero distances/dissimilarities), for all those coefficients identified above (as ratios). This is essentially anything with a y term or p_{12} in the denominator, since these give an Undefined! resemblance entry for blank samples. The pure distance measures D_1, D_6, D_7 and D_{10} will be unchanged with an added dummy, as will the species-standardised S_{15} (which promptly has to remove the just-added dummy variable since its range R_i over samples is zero!)]

c) On Ekofisk macrofauna counts take **Analyse>Resemblance>**(Measure•Other:S14 Ochiai(P/A)), renaming the result to Ochiai (P/A).

Unravelling resemblances

v7

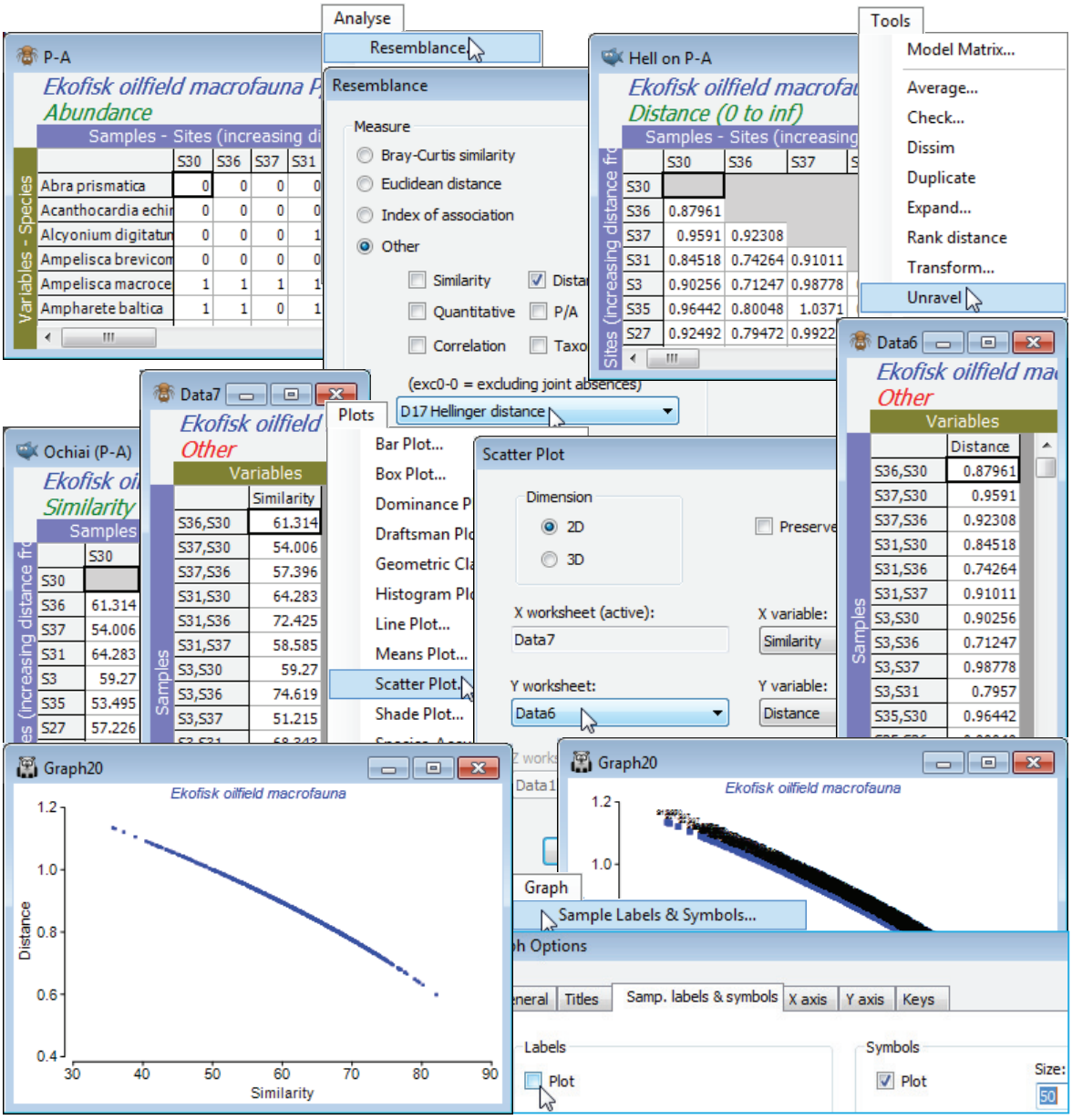
To view the relationship between these matrices, exploit two of the new features in PRIMER 7:

d) Run **Tools>Unravel** on both Hell on P-A and Ochiai (P/A), to turn these triangular matrices into long single columns (unravelling the rows), possibly now called Data6 and Data7.

Scatter plots

v7

e) With Data7 (say) as the active sheet, take **Plots>Scatter Plot**>(Dimension•2D) & (X variable: Similarity) & (Y worksheet: Data6) & (Y variable: Distance) – of course the X worksheet is the active Data7 – to see that Hellinger distance (on P/A data) is a decreasing function (near-linear here) of Ochiai similarity. The unnecessary sample labels can be removed by **Graph>Sample Labels & Symbols**, unchecking Labels✓Plot, and perhaps reducing the Symbols to Size: 50.



Other coefficients

Returning to the quantitative resemblance coefficients in the •Others list, five further measures given under the ✓Distance/dissimilarity heading are (loosely) based on likelihood-ratio tests. All are motivated by the (usually unrealistic) model in which the individuals of a species are randomly distributed in space or time (i.e. the data are strict counts, Poisson distributed), independently of other species, and with the mean count differing over species. A generalised likelihood ratio (GLR) test that two samples come from the same assemblage then produces the test statistic:

$$D^{BinD} = 2.\sum_i \left[y_{i1} \log\left(\frac{y_{i1}}{y_{i1}+y_{i2}} \right) + y_{i2} \log\left(\frac{y_{i2}}{y_{i1}+y_{i2}} \right) + (y_{i1}+y_{i2}) \log 2 \right]$$ Binomial deviance,

where the sum is over all p species as usual (note the first two terms do go to zero, unambiguously, when y_{i1} and y_{i2} are zero, respectively). In fact, the coefficient is of the form $2\Sigma[O\log(O/E)]$, where $O=y_{i1}$ or y_{i2} and $E=(y_{i1}+y_{i2})/2$ are the observed and expected values in a chi-squared type test of equality of counts for species i, then summed over the (supposedly independent) species, $i = 1,..., p$. The more familiar Wald test statistic for this situation is $\Sigma[(O - E)^2/E]$, but the two measures are likely to behave very similarly in practice (both having large-sample distributions of χ^2 on p df). A more useful variant of the latter is therefore given under Measure•Others, by simply dividing the chi-squared by the number of non jointly-absent species (p_{12}) for these two samples:

$$D^{Wald} = \frac{1}{p_{12}} \sum_i \left[\frac{(y_{i1}-y_{i2})^2}{(y_{i1}+y_{i2})} \right]$$ Wald (chi-squared) coefficient,

thus making this form of the coefficient independent of joint absences. This could be further modified in a natural way, to make it more robust to large y_{ij} (outliers) whilst preserving similar behaviour, by replacing a sum of squares with a sum of absolute values:

$$D^{Chi} = \frac{1}{p_{12}} \sum_i \left[\frac{|y_{i1}-y_{i2}|}{\sqrt{y_{i1}+y_{i2}}} \right]$$ 'Chi' statistic.

All three coefficients above are not dimensionless, i.e. they make sense only when applied to real counts and not densities, biomass, area cover etc. Millar RB & Anderson MJ 2004, *J Exp Mar Biol Ecol* 305: 191-221 therefore suggest a scale-invariant form of the first one:

$$D^{SBinD} = \sum_i \frac{1}{(y_{i1}+y_{i2})} \left[y_{i1} \log\left(\frac{y_{i1}}{y_{i1}+y_{i2}} \right) + y_{i2} \log\left(\frac{y_{i2}}{y_{i1}+y_{i2}} \right) + (y_{i1}+y_{i2}) \log 2 \right]$$

Binomial deviance (scaled).

(They choose to drop the 2 outside the sum and work in logs to the base 10, so for consistency with that paper, PRIMER does the same. Resulting analyses would be unchanged either way, since the difference is just the same constant multiplier for all pairs of samples). Because of the close link between likelihood ratio and Wald statistics, D^{SBinD} is seen to be a form of Clark's divergence, D_{11}, though without the adjustment for double zeros that comes through the p_{12} divisor.

Cao Y, Bark AW, Williams WP 1997, *Hydrobiologia* 347: 25-40 suggested a coefficient which has been advocated or used in subsequent studies. It looks very reminiscent of the (scaled) likelihood ratio statistic, but with an important switch of the y_{i1} and y_{i2} inside the logs:

$$D^{CY} = -\frac{1}{p_{12}} \sum_i \frac{1}{(y_{i1}+y_{i2})} \left[y_{i1} \log\left(\frac{y_{i2}}{y_{i1}+y_{i2}} \right) + y_{i2} \log\left(\frac{y_{i1}}{y_{i1}+y_{i2}} \right) + (y_{i1}+y_{i2}) \log 2 \right]$$ CY.

(It does take positive values in spite of the negative sign outside the sum!). Like D^{Wald} and D^{Chi}, it too contains the important p_{12} denominator adjustment to ignore joint absences, which the binomial deviance measures omit, but like D^{SBinD} it adds a denominator scaling to make the measure scale-invariant. However, it is now undefined when either $y_{i1} = 0$ (and $y_{i2} \neq 0$) or vice-versa, which could be much of the time, in fact! Zeros have to be replaced with a small positive number therefore, and the outcome is sensitive to this choice. No theoretical basis has been advanced for this coefficient, and it does not have an intuitively simple form, so any good operational properties it may possess must be somewhat fortuitous, and it is probably best avoided by the novice user.

Between-curve distances

Another useful application of multivariate methods was touched on at the end of Section 4, namely the analysis of structured sets of curves or (pseudo-)frequency distributions, generically referred to as sample *profiles*. These include particle- or body-size analyses, or growth curves, with several replicate profiles from each of a number of sites, times, treatments etc. Simple univariate statistical treatment of the size variable is often impossible because of the inherent serial correlation problems (*repeated measures*) of, for example, tracking the body size of a single organism through time, or the lack of a proper frequency distribution structure in histograms of particle sizes (in no sense are we counting independent particles entering the sampling device, to give multinomial frequencies). A viable multivariate alternative is to treat the independent units as the whole profiles and define distances among them, taking these pairwise resemblances into, say, the ANOSIM tests discussed in Section 9. Suitable distance measures between pairs of curves include Euclidean distance D_1 (or its square), the Manhattan distance D_7 and, specifically for comparing cumulative curves:

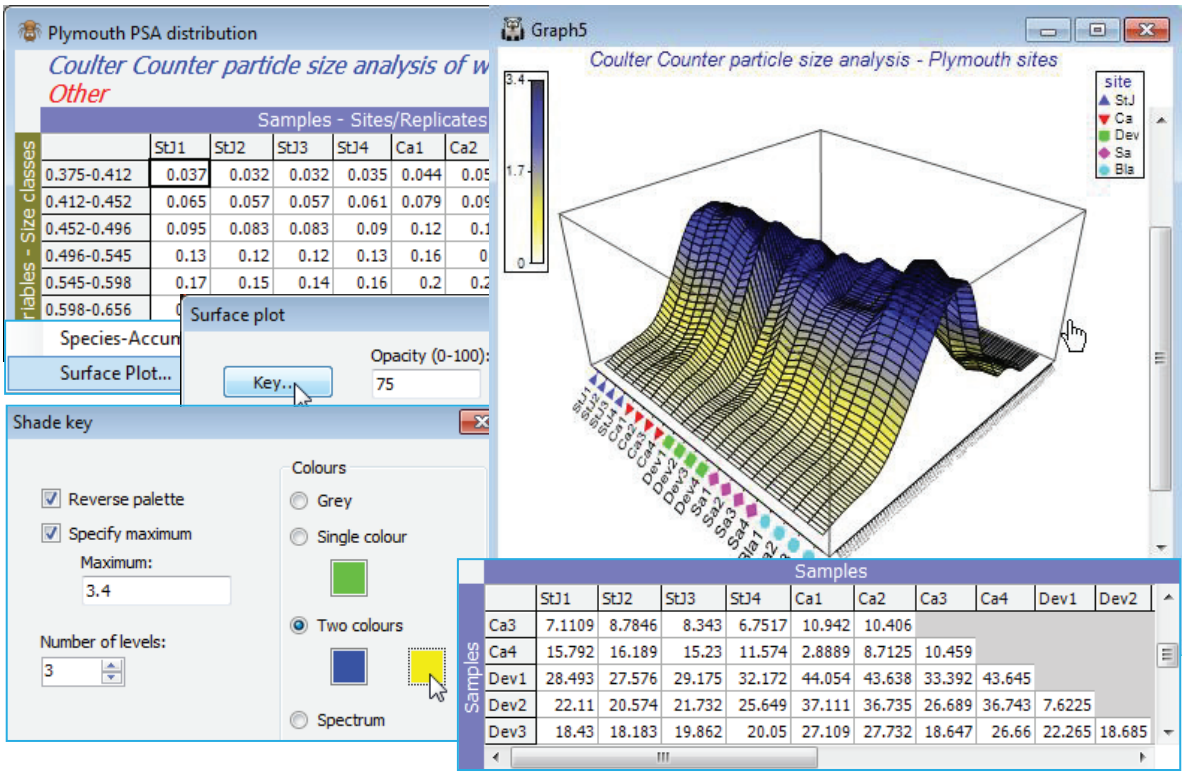
$$D^{\max} = \max_i |y_{i1} - y_{i2}| \qquad \text{Maximum distance,}$$

which is also a Distance/dissimilarity option on the •Others list. The maximum departure of two cumulative frequency curves from each other, taken over all the size categories, is the basis of the Kolmogorov-Smirnov test, but the testing structure there relies on real (multinomial) frequencies. Where this is not the case, as often, maximum departure may still be a sensible distance measure of two curves to feed into multivariate analysis, though Manhattan (or Euclidean) distance is likely to be at least as good, since it sums positive contributions across the entire size range.

(Plymouth particle-size analysis)

An example of a particle-size analysis (PSA) matrix has already been seen for Danish sediments at the end of Section 4, for which the histogram was smoothed by cumulating the size-classes. Here we examine instead an already smooth frequency distribution from Coulter Counter processing of water samples, in which large numbers of suspended particulates are automatically sized into one of 92 logarithmically increasing particle-diameter ranges (the variables). The samples are of four replicate water samples from each of five Plymouth sites, and some analysis is presented towards the end of Chapter 8 of CiMC. The directory C:\Examples v7\Plymouth PSA holds the frequency distributions in Plymouth PSA distribution. Columns are samples, and entries are % particles in each size-class (add to 100). The curves are conveniently viewed with **Plots>Surface Plot**, colour changed with **Graph>Special>Key**, zoomed with **Graph>Zoom In** or the icon, and rotated with **Graph>Rotate Axes** or the icon (hold-click and move cursor). Create Manhattan distances among the curves, with **Analyse> Resemblance>**(Measure•Other: D7 Manhattan distance), which could go into ANOSIM to test for characteristic site differences in PSA profiles at these times.

v7



Taxonomic distinctness/ aggregation files

A later section (15) discusses univariate diversity indices that can be computed from each sample, including biodiversity measures that are based on the relatedness of the species making up a simple species list (P/A data), see Chapter 17 of CiMC. Though the supplied relatedness could be genetic, phylogenetic or functional – through suitable provision of a distance/dissimilarity matrix among the species, perhaps (but not necessarily) their pairwise distances apart through some hierarchical arrangement of species – PRIMER 7 implements the idea mainly in terms of *taxonomic distinctness* (see Section 15). These are the distances travelled in connecting every pair of species through a tree with a fixed set of levels (typically, a Linnaean taxonomy). If, on average, these distances are large, then the sample is considered biodiverse. A necessary input is a *variable information* sheet, which (for historic reasons) PRIMER calls an *aggregation* file (see the end of Section 2), defining the taxonomy – which species belong to which genera, families, orders, etc. From this, path weights ω_{ij} are calculated between every pair of species, i and j. Always, ω_{ij} takes the value 100 for two species that are connected at the most distant level; e.g. if the final column heading in the taxonomy file is *phylum* then two species in different phyla are defined to be 100 units apart (do not add a final column, say *kingdom*, for which all species have the same entry, Animalia; you could then only attain the value 100 for species in different kingdoms). By default, intervening levels are considered to be equally-spaced. For example, for a hierarchy of species from different classes all in the same phylum, with the five levels of species, genus, family, order and class, two species in the same genera are 20 units apart, in different genera but the same family are 40 units apart, etc. This can be overruled in two ways: either a user can define his/her own step branch-lengths, which will again be rescaled to a maximum of 100 for two species in different top-level groups, whatever scale is input for the absolute steps; or the information in the aggregation matrix about taxon richness at each hierarchical level can be used (a level in the tree which has almost as many taxa as the level below it gives rise to a step of shorter branch-length).

Taxonomic dissimilarity measures

This concept of taxonomic distinctness can be carried over from a diversity index to a dissimilarity coefficient. Two measures are given under **Analyse>Resemblance**>(Measure•Other: ✓Taxonomic P/A). Both are presence/absence measures only, indicated by the plus sign superscript: Γ^+ (upper case Greek gamma) is a natural extension of Bray-Curtis dissimilarity on P/A data (the latter is just the complement of Sørensen S_8), and Θ^+ (upper case Greek theta) similarly extends Kulczynski P/A dissimilarity, the complement of S_{13}. They are formally defined as:

$$\Gamma^+ = \frac{\left(\sum_{i=1}^{s_1} \min_j \{\omega_{ij}\} + \sum_{j=1}^{s_2} \min_i \{\omega_{ij}\}\right)}{(s_1 + s_2)}, \qquad \Theta^+ = \frac{1}{2}\left(\frac{\sum_{i=1}^{s_1} \min_j \{\omega_{ij}\}}{s_1} + \frac{\sum_{j=1}^{s_2} \min_i \{\omega_{ij}\}}{s_2}\right)$$

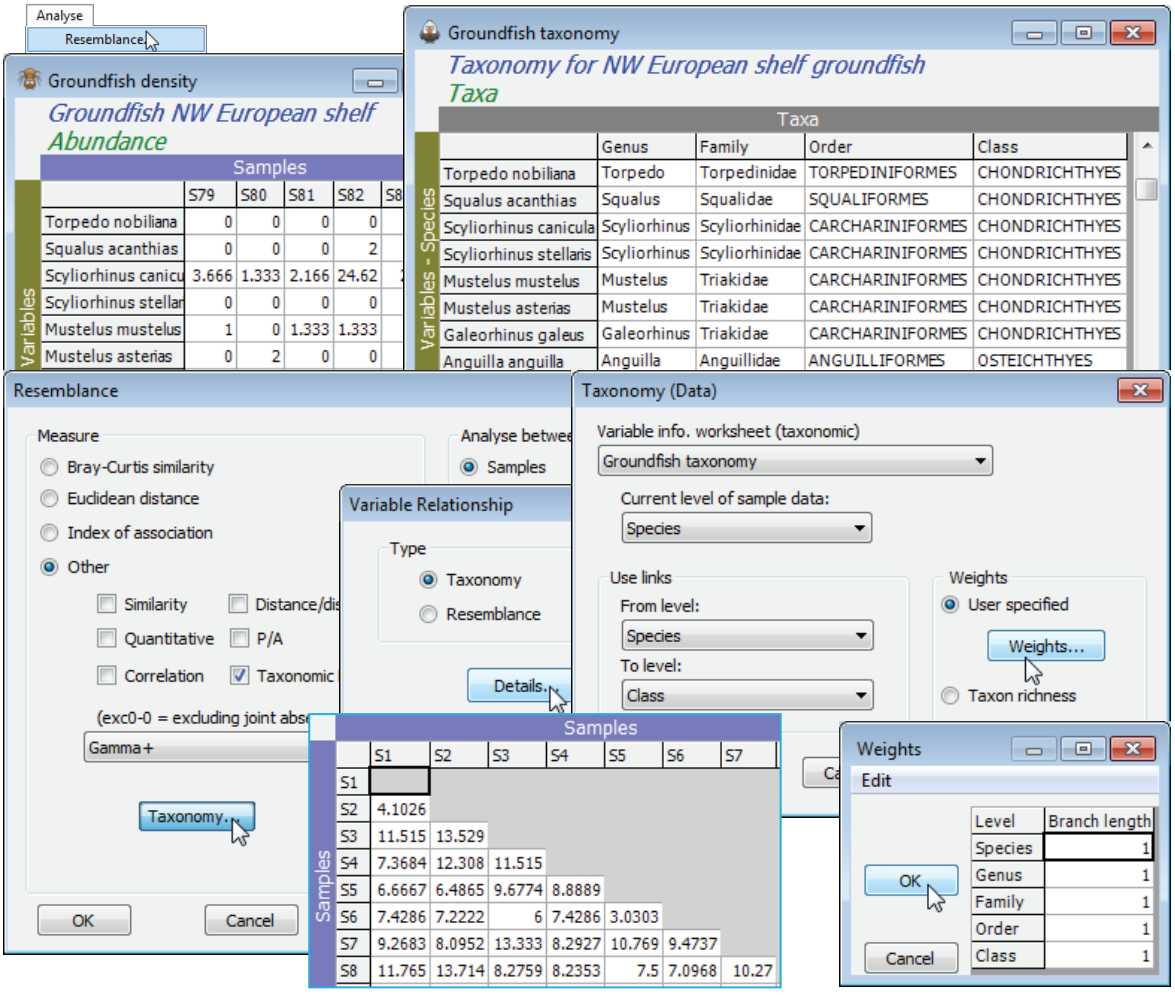
where there are s_1 species present in sample 1 and s_2 in sample 2, and ω_{ij} is the distance through the tree from species i of sample 1 (i = 1, 2, …, s_1) to species j of sample 2 (j = 1, 2, …, s_2). This is almost simpler to express in words: for each species one finds the most closely related species in the opposite sample, then averages these minimum path lengths over all (s_1 + s_2) species, to obtain Γ^+. (If the nearest relation in the opposite sample is the same species, the path length is defined to be zero, of course). For Θ^+, these averages are calculated separately, i.e. the average path length for all species in sample 1 to their nearest neighbours in sample 2, then for all species in sample 2 to their nearest neighbour in sample 1, with these two averages then themselves being averaged.

As noted, these constructions result in Γ^+ and Θ^+ reducing to the dissimilarity forms of Sørensen and Kulczynski (P/A) when the hierarchy collapses, i.e. when all species are in one higher-order group and the path lengths are 0 or 100 (species do or do not have a match in the opposite sample).

Θ^+ was defined (and referred to as an 'optimal mapping statistic', denoted M) by Clarke KR & Warwick RM 1998, *Oecologia* 113: 278-289, and Γ^+ is (to within a constant) the TD of Izsak C & Price ARG 2001, *Mar Ecol Prog Ser* 215: 69-77. They are clearly closely related, and will be identical when $s_1 = s_2$. Their use is in ordinating samples from widely-spread biogeographic regions with few, if any, shared species, but which will always have higher-order taxa in common. They also provide a certain amount of robustness in dissimilarity value to mistakes or inconsistent identification at the finest taxonomic levels (see CiMC, end of Chapter 17, for two applications from Clarke KR, Somerfield PJ, Chapman MG 2006, *J Exp Mar Biol Ecol* 330: 55-80).

(Groundfish of European shelf waters)

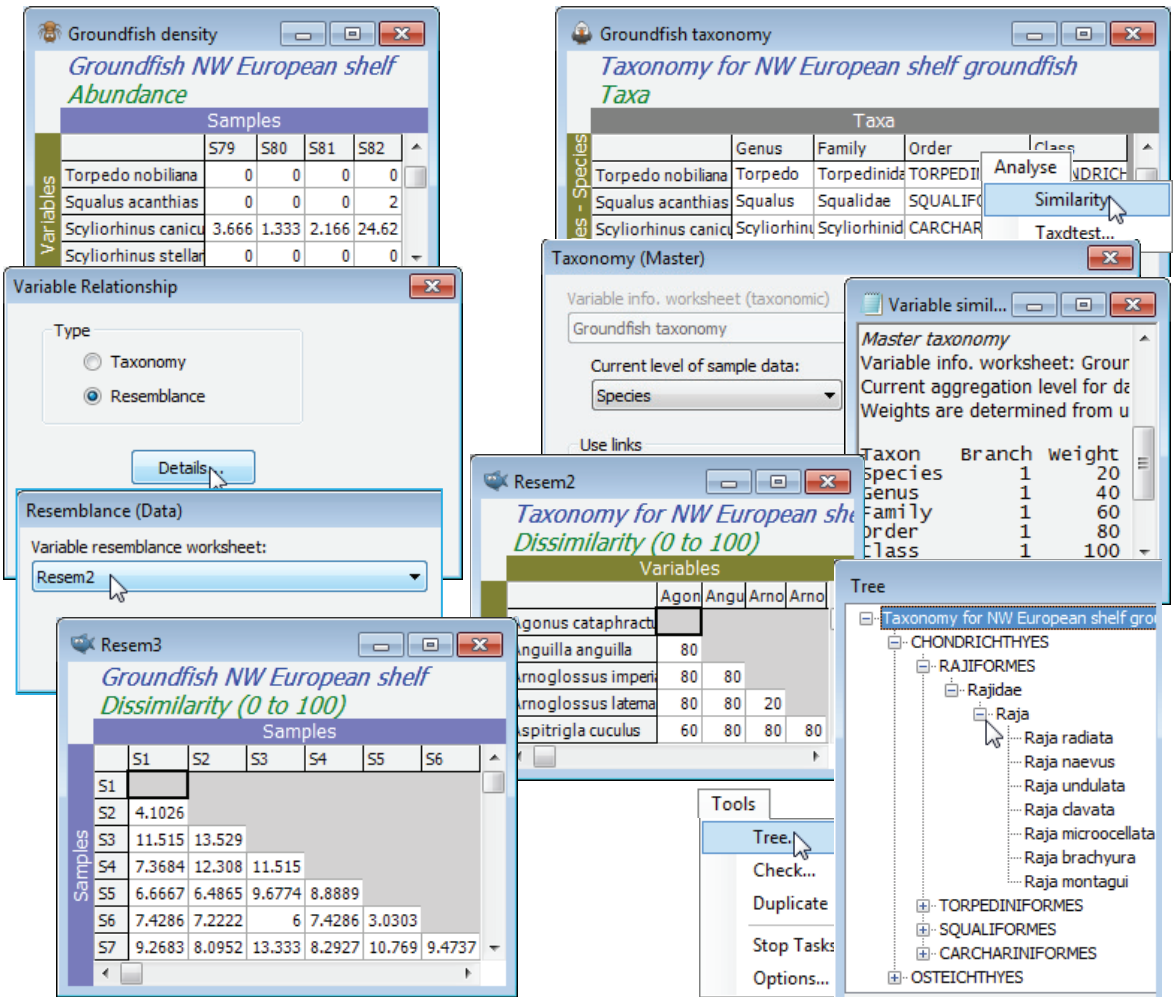
Assemblage data from 93 groundfish species, those that could be reliably sampled and identified in beam-trawl surveys by research vessels from several countries surrounding NW European shelf waters, were analysed by Rogers SI, Clarke KR, Reynolds JD 1999, *J Anim Ecol* 68: 769-782. The data matrix, C:\Examples v7\Europe groundfish\Groundfish density(.pri) is of 277 locations (ICES quarter-rectangles) sampled in the third quarter of the year over the period 1990-96, with the values being mean catch rates corrected to number of fish per 8m beam trawl per hour. The sites are divided *a priori* into 9 coastal areas (1 to 9 in factor area: 1=Bristol Channel, 2=Western Irish Sea, ..., 9=E Central North Sea, see the **Edit>Properties** box). Also available is a PRIMER format variable information file, the *aggregation* sheet, Groundfish taxonomy(.agg), met briefly at the end of Section 2. Open both into a new workspace and on Groundfish density **Analyse>Resemblance>** (Analyse between•Samples)&(Measure•Other: Gamma+>**Taxonomy**>(Type•Taxonomy)>**Details** >(Variable info. worksheet (taxonomic): Groundfish taxonomy)), accepting all defaults on this last dialog box – though you might like to take User specified>**Weights** to note how the step lengths between levels are set to be equal, resulting in path lengths between species of 0, 20, 40, 60, 80, 100, as earlier described. Alternatives might be to flatten the tree at the top, by setting the final step (Class) to 0, or to give more weight to the fine-level taxonomy, with decreasing entries 5, 4, 3, 2, 1.



Relatedness supplied as resemblances

v7

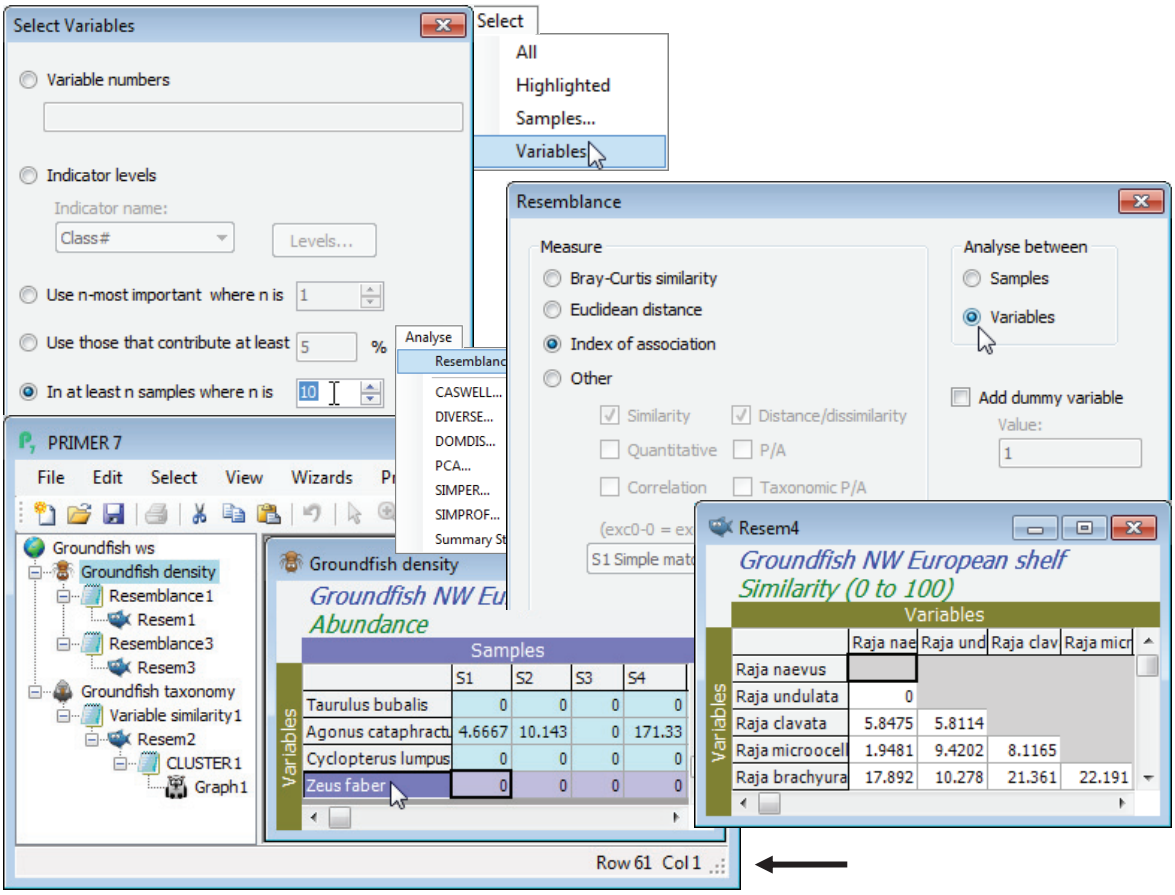
Note the alternative means of supplying the variable information, to these dissimilarity measures and the biodiversity indices of Section 15, which is now available in PRIMER 7. In the Variable Relationship dialog box, Type•Resemblance>**Details** now requires specification of a numeric among-species resemblance matrix which could be constructed from genetic, functional etc data, but is illustrated here by first creating a species distance matrix through the Linnean tree with **Analyse>Similarity** when the aggregation file Groundfish taxonomy is the active window. This takes you to a similar Taxonomy dialog box as above and creates sheet *Resem2* of among-species distances 20, 40, 60, 80, 100. The Linnean tree could be viewed by **Analyse>Cluster** (next section) on Resem2, or in alternative format by **Tools>Tree** on Groundfish taxonomy. When Resem2 is supplied as the Variable resemblance worksheet from **Details**, the same Γ^+ matrix results, naturally.



Analysing between variables

The introduction above of the concept of 'distances' among species raises the issue of how best to compute *species similarities* – or more generally *variable associations* – taking the menu option of **Analyse>Resemblance>**(Analyse between•Variables). Several significant new developments in PRIMER 7 (see Section 10 and Chapter 7 of CiMC) on *shade plots* and *coherent species curves* concern better display and analysis techniques for characterising responses of individual (or groups of) species across the samples in space, time or over a changing environment. Two species are considered perfectly similar if they co-occur across samples – with numbers or biomass in strict proportion, for quantitative data. As with sample similarities, the issue of how to treat joint absence is often relevant here too – it would often be appropriate to regard the absence of two species at a particular site as uninformative (a clay-living and a gravel-living species are not similar because neither are found at sandy sites). A measure which captures this biological constraint well is the Whittaker Index of Association (IA), met earlier in this section. As remarked then, this will give the same outcome as standardising species (by total) and applying Bray-Curtis on the species. The (implicit) standardisation however will come unstuck with all-blank species, which must certainly be removed, and it is also almost always a good idea to remove all the 'occasional' species, rarely observed and with low abundances when they do occur. The various options for reducing to the 'most important' species were covered at the end of Section 3, and for standardising species near the start of Section 4; these options would not usually apply when calculating sample similarities, but are important to eliminate wildly erratic, and not meaningful, similarities among rare species.

On the Groundfish density matrix, **Select>Variables>**(•In at least n samples where n is: 10). To see how many species are retained (61, in fact), click on the sheet's final row which displays this in the status bar at the foot of the PRIMER desktop. Take **Analyse>Resemblance**>(Measure•Index of association) & (Analyse between•Variables) to create the species similarities (*Resem4* perhaps). Show that the same outcome is produced (*Resem5*) by putting the selected species from Groundfish density through **Pre-treatment>Standardise**>(Standardise•Variables) & (By•Total), followed by **Analyse>Resemblance>**(Measure•Bray-Curtis similarity) & (Analyse between•Variables).



Correlation between variables

One context in which resemblances between variables is often of primary interest is in dealing with environmental variables, biomarkers, morphology etc. Concepts of ignoring joint absences do not apply – in fact zero no longer necessarily means absence (e.g. 0°C), particularly after normalisation (see Section 4). Variables are usually on different measurement scales (or are non-comparable on the same units), so correlation is a natural choice, with its built-in normalisation. The final option in Measure•Others is ✓Correlation, with seven variations of a correlation coefficient, ρ, namely

$$\rho^P = \frac{\sum_j (y_{1j} - y_{1\bullet})(y_{2j} - y_{2\bullet})}{\sqrt{\sum_j (y_{1j} - y_{1\bullet})^2 \sum_j (y_{2j} - y_{2\bullet})^2}} \qquad \text{Pearson (product-moment) correlation,}$$

where $y_{1\bullet} = (\sum_j y_{1j})/n$ is the average of the n sample readings for variable 1, etc, and two non-parametric choices, based only on rank values (r_{ij}), the numbers 1, 2, 3, .., n across samples j, for each variable i. Spearman is simply Pearson correlation calculated on the ranks, reducing to:

$$\rho^S = 1 - \frac{6}{n(n^2-1)} \sum_j (r_{1j} - r_{2j})^2 \qquad \text{Spearman rank correlation,}$$

v7 and Kendall's τ is an alternative (Kendall MG 1970, *Rank correlation methods*, Griffin, London), which in practice tracks Spearman closely, but with lower absolute values. These three coefficients are then given as absolute values, $|\rho|$, to cater for situations where it is not especially meaningful to distinguish between positive and negative correlations (e.g. some biomarkers increase under impact and some decrease, so an absolute ρ is often a better description of their inter-relationship). A final weighted form of Spearman gives more emphasis to small ranks (high variable values):

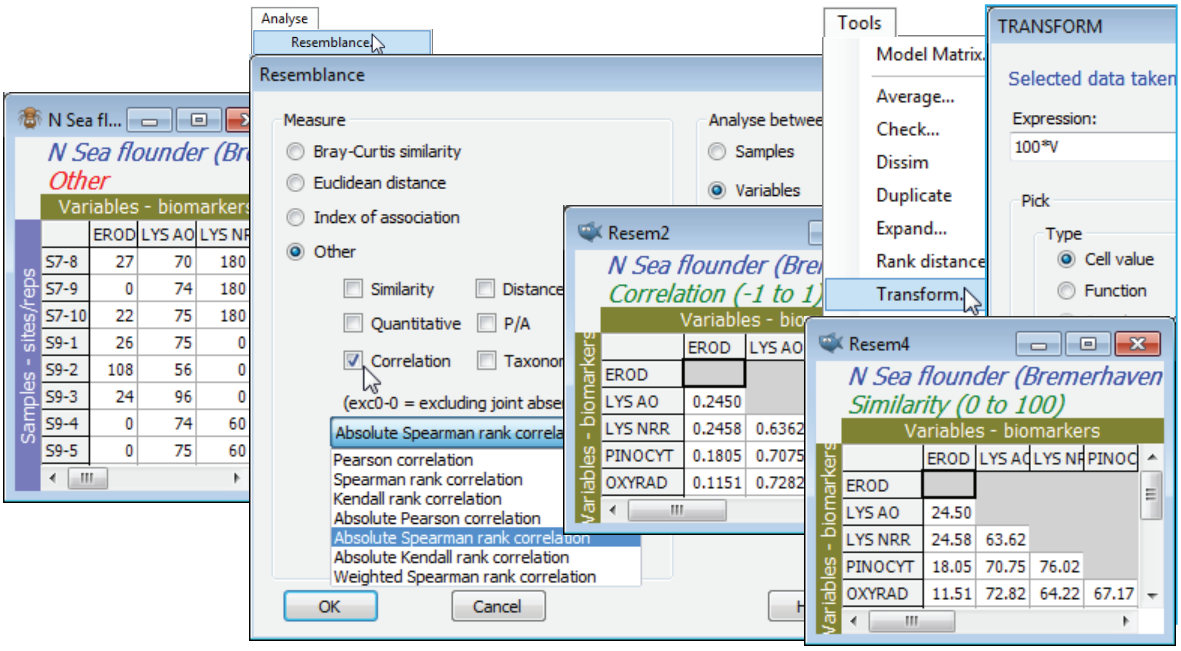
$$\rho^W = 1 - \frac{6}{n(n-1)} \sum_j \frac{(r_{1j} - r_{2j})^2}{r_{1j} + r_{2j}} \qquad \text{Weighted Spearman rank correlation,}$$

but this really only makes sense in an asymmetric context, such as correlating the entries of two resemblance matrices, thus emphasising matching pairs of high similarities – see the discussion of equation (11.4) in CiMC.

Correlation as similarity

Use of a correlation matrix between all pairs of variables as input to a multivariate ordination (say), in which points denote variables rather than samples (so that highly correlated variables are placed close together), either requires one of the absolute coefficients or a simple shift $S = 50(1+\rho)$ of the three standard coefficients, so that they are defined over (0, 100) rather than (–1, 1). There is an important difference between the two approaches: should highly negatively correlated variables be considered highly similar (use an absolute measure) or highly dissimilar (shift the scale upwards)? The practical context should usually make clear which is the right choice.

Save and close the current Europe groundfish workspace (as Groundfish ws), and open that for the N Sea biomarkers N Sea ws, created towards the end of Section 4 – see there for description of the variables. (If not available, just open N Sea flounder biomarkers(.pri) from directory C:\Examples v7\N Sea biomarkers). The previous pre-treatment by *variability weighting* of these (transformed) biomarkers was designed for calculation of standard sample similarities (which you may now wish to do by **Analyse>Resemblance>**(Measure•Euclidean distance) & (Analyse between•Samples)), but the reason for re-opening this workspace now is to calculate similarities among variables, via correlation. The choice is between standard (Pearson) correlation and a rank-based correlation (Spearman, say); if the analysis includes the categorical as well as the continuous variables, the rank option may be preferred. Note that any variability weighting previously carried out, to weight the biomarkers against each other in calculating sample similarities, will be irrelevant to correlation computation of variable similarities, because variables are renormalised (under Pearson) or ranked (under Spearman). For Spearman, even the square root transform applied to the EROD and Lipid variables is irrelevant, since this will not change the rank order of variable values across samples. Note that low lysosomal stability (AO or NRR) is associated with high EROD etc – both indicating contaminant impact – so an absolute correlation measure is used to capture biomarker similarities. **Analyse>Resemblance>**(Measure•Other>✓Correlation: Absolute Spearman rank correlation) & (Analyse between•Variables) on N Sea flounder biomarkers will produce values in the range (0,1). These could be scaled to (0,100) using **Tools>Transform>**(Expression:100*V) – see box heading Transform on resemblances in Section 11 – and the Type changed from Correlation to Similarity with **Edit>Properties>**(Resemblance type•Similarity) but this is not practically necessary for most routines in PRIMER, such as nMDS ordination, since only ranks of the resemblances are used.



Corrections for missing data

v7

Returning to the main purpose of resemblance measures, to describe similarity among samples, an important new feature in PRIMER 7, not offered in earlier versions, is that resemblance measures will now be calculated in the presence of missing cells (identified by Missing! in the sheet). As described in Section 1 (box heading Missing or zero values?) this tends to arise only for sheets of type Environmental or Other – species matrices can have whole samples missing from an otherwise balanced layout but this is not regarded as missing data, just unbalanced design, handled routinely in PRIMER (and PERMANOVA+). Under restrictive conditions (multivariate normality in a 'not

too high' dimensional space) it may be possible for some environmental data to estimate single entries missing at random, utilising the correlations between variables (see **Tools>Missing** in Section 12) but in many contexts for which missing entries are almost guaranteed, these modelling conditions will not apply. An example would be would be questionnaire data, in which the samples are the individual respondents and the variables the questions, e.g. with matrix entries 1 to 5, for a 'disagree strongly' to 'agree strongly' scale. This is a likely area for application of multivariate methods, calculating similarities between respondents in the profile of answers, and linking this to demographic/socio-economic data, e.g. PRIMER applications from environmental economics exist, but missing answers are commonplace and probably not estimable under normality assumptions.

v7

Where there is missing data, PRIMER 7 therefore computes a resemblance between each specific pair of samples by removing (for that calculation only) those variables in which one or other value is missing (referred to as *pairwise elimination* of missing data). This can cause a crude bias in some distance measures which are in the form of sums rather than averages of variable contributions, in that pairs of samples with many missing entries will automatically return lower distances than those with few or no missing values, all else being equal. Examples are Euclidean (D_1) or Manhattan (D_7) distances, which are both based on simple sums over the variables. A correction for these biases is straightforward in this case: average Euclidean distance (D_2) clearly has no such crude bias since the contributions from each variable are averaged not summed. The solution for D_1 is therefore to multiply up the summation by a factor (p/p'), where p is the full number of variables in the array, and p' is the number of variables used in that specific sum, having pairwise-eliminated the missing variables. The outer square root in the definition of D_1 makes the overall correction term $(p/p')^{0.5}$.

PRIMER 7 automatically applies such correction factors to every resemblance measure, if needed, as shown in the following table. Note that the standardisation implicit in many measures, including all (dis)similarities, avoids the need for correction, sample totals always being re-defined for each pairwise-eliminated set. The corrections have only asymptotic justification for the more complex measures, e.g. D_{16} Chisquared distance for which the correction term is $(p'/p)^{0.5}$, not $(p/p')^{0.5}$, thus a downward adjustment. (Similarly, that for Maximum Distance is based on Jensen inequalities on asymptotics of extreme value distributions so is definitely approximate!). It should be stressed that these corrections assume an average contribution from each missing variable, as measured by the average for the present variables. Broadly, this is not unreasonable if values are missing at random, but is theoretically inferior to reconstruction of missing values by **Tools>Missing**, when the strict conditions for this apply, since that uses variable correlations to estimate non-average values.

Distance/dissimilarity (quantitative, + P/A)	
D_1 - Euclidean	$(p/p')^{0.5}$
D_2 – Average Euclidean	None
D_3 - Chord	None
D_4 - Geodesic	None
D_6 - Minkowski	$(p/p')^{1/r}$
D_7 - Manhattan	p/p'
D_8 - Czekanowski (exc0-0)	None
D_{10} - Canberra metric	p/p'
D_{11} - Divergence (exc0-0)	None
D_{13} - Non metric coeff +	None
D_{14} - Bray-Curtis dissimlty	None
D_{15} – Chisqrd metric	$(p'/p)^{0.5}$
D_{16} – Chisqrd distance	$(p'/p)^{0.5}$
D_{17} - Hellinger	None
Gamma +	None
Theta +	None
CY	None
Binomial deviance (scaled)	p/p'
Binomial deviance	p/p'
Wald test (chisquared)	None
Chi statistic	None
Maximum distance	$[\log(p)/\log(p')]^{0.5}$
Modified Gower	None

Similarity (P/A)	
S_1 - Simple matching	None
S_2 - Rogers & Tanimoto	None
S_5	None
S_6	None
S_7 - Jaccard	None
S_8 - Sørensen	None
S_{11} - Russel & Rao	None
S_{13} - Kulczynski P/A	None
S_{14} - Ochiai P/A	None
S_{26} - Faith	None
Similarity (quantitative)	
S_{15} - Gower	None
S_{17} - Bray-Curtis similarity	None
S_{18} - Kulczynski (quant)	None
S_{19} - Gower (exc0-0)	None
Canberra similarity (exc0-0)	None
Ochiai similarity (quant)	None
Index of Association	None
Correlation	
Pearson correlation	None
Spearman correlation	None
Kendall correlation	None
Weighted Spearman	None

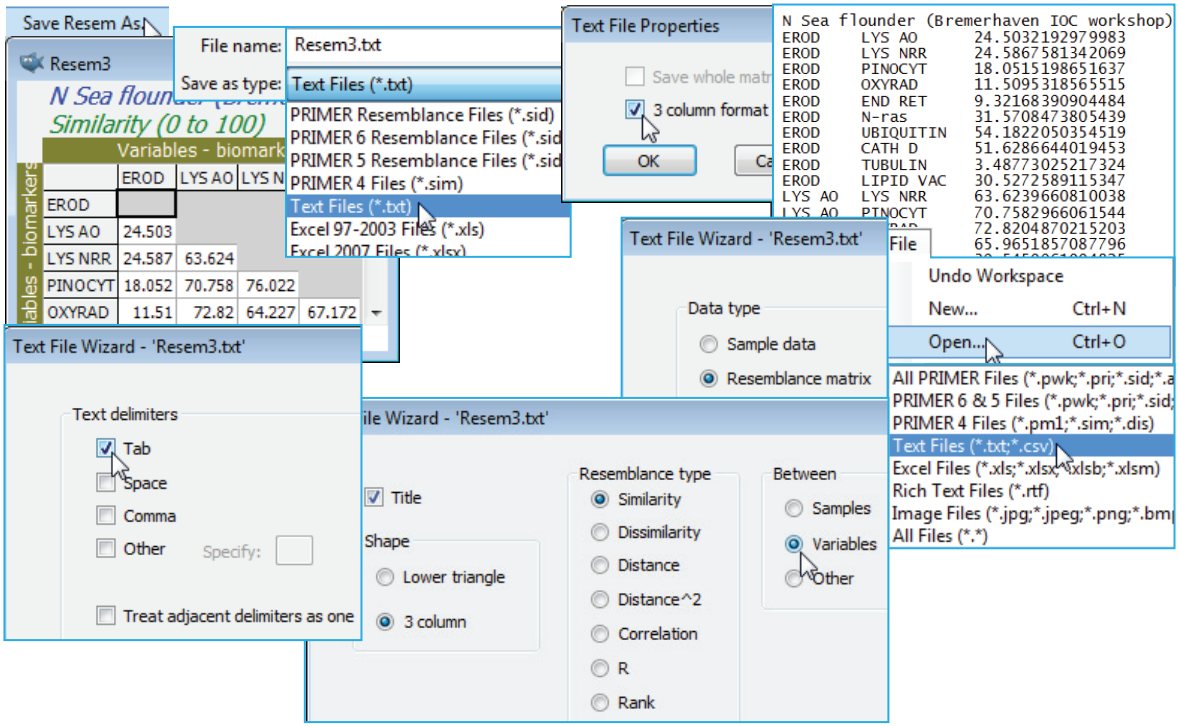
Saving & opening triangular matrices

File>Save Resem As will save a resemblance matrix in internal binary PRIMER v7 (*.sid) format, though the previous v6 and v5 binary formats (also *.sid) are other options – as is the early DOS text format (*.sim) – all likely to be of limited utility now. More useful are the options to save the triangular matrix as an Excel sheet (*.xls or *.xlsx), in which case the diagonal and upper triangular cells are left blank. Several text file choices (*.txt) are also offered: by default a lower triangular matrix is output with tabs as separators, though there is also the option to output a 'whole matrix', i.e. a full square is saved, with filled diagonals and upper triangle as the transpose of the lower half. Another interesting possibility is a 3-column output format, with first and second columns giving the row and column labels for the lower triangle, and the third column the resemblance entry. (This parallels the 3-column – *flat-form* or *record format* – data files, the output or input of which was seen in Section 1). These options should, between them, make it easy to take a resemblance matrix out of PRIMER into other software, if needed.

v7

File>Open>(Data type•Resemblance matrix) gives all these options in reverse (and more), for reading in any triangular matrix. Generic questions concern the existence or otherwise of a ✓Title, and a type specification of: •Similarity/ •Dissimilarity /•Distance /•Distance^2 /•Correlation /•R /•Rank (the notation R come from pairwise ANOSIM statistics, see Section 9, but could represent any measure defined over (-1, 1) for which the larger the value the greater the 'distance apart'). Whether input matrices are to be treated as (Between•Samples) or (Between•Variables) is also required, of course. Excel files (*.xls or *. xlsx) are assumed to be in lower triangular form – if an upper triangle or diagonal is present it is ignored. Text files have more options, the choices being: (Shape•Lower triangle) or (Shape•3 column). Both of these lead to the same 'Text File Wizard' dialog seen in Section 1 for inputting data matrices, in which any form of separator between entries can be defined, even in combination. [Thus, though of limited usefulness, if unravelled distance matrices – as created for the scatter plot of (Hellinger on P/A) *vs.* (Ochiai P/A) earlier in this section – were saved as *.txt data files (of two columns), with care they could be read back into PRIMER to reform the triangular matrices. To do this, you would need to say you are inputting resemblances in 3-column format, and take both ✓Tab and ✓Comma as Text delimiters, allowing interpretation of the 1st column ('row label,col label') as columns 1 and 2 of the 3-column format.]

Try saving the previously created 'variable similarities' matrix among biomarkers (from the N Sea ws workspace which should still be open) into Excel and text formats, in both standard *.txt and the 3-column *.txt formats. Look at these in Word or Notepad, and then try re-opening them again in PRIMER. Resave the workspace, N Sea ws, for a later section and close it.



6. Clustering methods (*CLUSTER*, *SIMPROF, UNCTREE, kRCLUSTER*)

Clustering methods & choice of linkage

v7

PRIMER 7 now carries out a wider range of clustering methods than previously: a) hierarchical agglomerative clustering using one of four linkage methods – single, complete, group average (UPGMA) and flexible beta (a standard WPGMA extension); b) hierarchical (binary) divisive clustering – a new unconstrained form (*UNCTREE*) of the previously offered constrained binary divisive routine (*LINKTREE*, covered in Section 13); and c) a new flat-form, i.e. non-hierarchical, method (*kRCLUSTER*) which is a development of *k*-means clustering. Both the *UNCTREE* and *kRCLUSTER* algorithms are designed to fit with the non-parametric approach which is central to the PRIMER package, e.g. by optimising the *ANOSIM R* statistic (see Section 9) as a measure of group separation based only on the ranks of the resemblance matrix. These new (and old) clustering methods, all accessed by **Analyse>CLUSTER**, are described in detail in Chapter 3 of CiMC. For most methods the output is a dendrogram, i.e. tree diagram, displaying a hierarchical grouping of samples (or sometimes of species, see Section 10), with a divisive hierarchy being differentiated visually from an agglomerative one by a slight change in the way the final pairings are displayed. The main output of the non-hierarchical *kRCLUSTER* method(s) is simply a factor (or indicator) specifying the group to which each of the samples (or species) is allocated. All routines can be applied directly to any of the triangular matrices produced by the **Analyse>Resemblance** menu.

SIMPROF tests

v7

All of the clustering methods are able to exploit 'similarity profile' (SIMPROF) permutation tests, e.g. for stopping rules for divisive methods or choice of number of groups *k* in a 'flat' clustering. SIMPROF test sequences look for statistically significant evidence of structure in samples which are *a priori* unstructured (e.g. single samples from each of a number of sites). Under this option, tests are performed at every node of a completed dendrogram, whether constructed agglomeratively or divisively, starting from the top of the dendrogram (all points in a single group) and permitting interpretation of divisions below each node only if a SIMPROF test shows evidence of multivariate structure within that group. Test results are displayed by a colour convention on the dendrograms: samples connected by red lines are not significantly differentiated by SIMPROF, so that only the structure shown by black lines in a dendrogram should be interpreted. The test statistics themselves and their significance levels are given in the Results window indicated by the  con.

SIMPROF on large matrices

v7

The dendrogram itself is rapidly calculated, at least for the agglomerative methods, since no search procedure is involved, and it can thus be constructed for very large numbers of samples – but the SIMPROF routine is highly compute-intensive, given the typical number of permutations (default 999) and recomputations of the similarities which are necessary for each nodal test (CiMC, Chapter 3), and the potentially large number of nodes. PRIMER 7 now allows the option (the default, which would normally be taken) of dividing these calculations among the multiple processors constituting the core of modern PCs, but it is still unwise to take routinely the SIMPROF option with very large resemblance matrices. A selective form of SIMPROF applied to a single selection of samples, and which provides graphical output of the similarity profile, the spread of alternative profiles obtained under permutations of the data matrix and the null hypothesis distribution for that single test, can be found on the **Analyse>SIMPROF** menu, when the active sheet is a (selection of a) data matrix. A possible strategy for large arrays, which are clearly not going to complete all the nodal tests in a viable time, may then be to carry out the clustering, having turned off the ✓SIMPROF test box on the Cluster dialog, and then manually choose a series of nodes from the dendrogram, testing them one at a time by selecting their samples and carrying out individual **Analyse>SIMPROF** tests, thereby getting some idea of how much structure is potentially interpretable from the dendrogram.

Modifying plots in PRIMER

v7

Though PRIMER 7 does not attempt to replicate all the facilities available in graphics presentation software, there are a large number of graphics options available to modify dendrograms, many of them shared in a consistent interface for ordinations and other PRIMER plots. Note that the range of plots available now in PRIMER 7 has expanded greatly, with a new **Plots** menu and the concept of a multi-plot. This are discussed in Section 7 but examples can be seen scattered throughout this User Manual/Tutorial where they are relevant to a particular analytical technique. Some general features are that plots can be: resized; titles, sub-titles and axis titles edited, and font colours, sizes and types changed; these and the history display removed altogether; sample and variable labels replaced with factor levels and/or symbols, the latter having a choice of symbol size, type and colour; lines thickened (unselectively); axes rescaled to specification or even logged, though this

v7

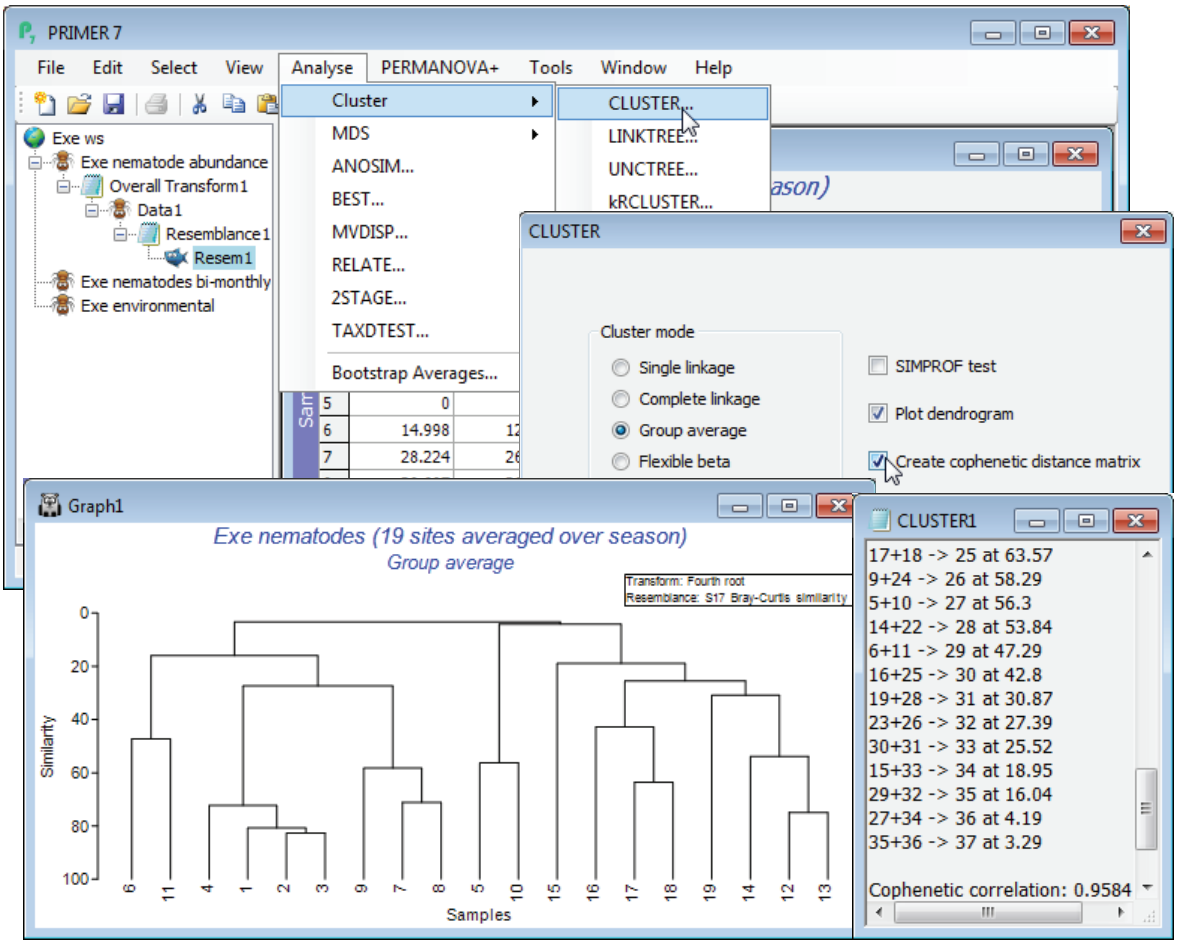
will not usually be appropriate for a dendrogram, etc. New general features in PRIMER 7 include the ability with a single check box to change colour symbols or lines to monochrome and colour shading to mono hatching patterns, and much better control in plot keys of size of symbols and size, font and title font of text. Features specific to dendrograms include: the ability to orient the plot in any of the four directions; to display a slice through the tree at a fixed resemblance level, and create a factor (/indicator) that defines the groups at that threshold; to rotate sub-groups of the tree, in any permissible way; and to collapse the detail of specific sub-groups, so that the overall structure of a large tree can be better displayed. The fine detail is seen by another general facility for all plots in PRIMER: a flexible zoom operation which maintains the position of labelling and axis scaling while zooming in on the content of the plot. Importantly for dendrograms, the aspect ratio of the zoomed box can also be changed, allowing clear presentation of detailed structure (this latter feature does not operate with ordinations, for good reason – see later!).

(Exe estuary nematodes)

Assemblage data on 140 species of free-living marine nematodes at 19 sites (labelled 1-19) in the inter-tidal soft sediments of the Exe estuary, UK, is in data file C:\Examples v7\Exe nematodes\ Exe nematode abundance(.pri); the entries are averaged counts over 6 bi-monthly samples in one year. An analysis of the full data, Exe nematodes bi-monthly(.pri) suggests that seasonality must be relatively weak, if present – see CiMC Fig. 6.12 – and this example is mainly used here, and in CiMC, in its time-averaged form. The file Exe environmental(.pri) contains six environmental variables for the sediments at those 19 sites: median particle diameter, depth of the water table, depth of the anoxic layer, height up the shore, % organics and interstitial salinity. The field study is described in Warwick RM 1971, *J Mar Biol Assoc UK* 51: 439-454 and the original multivariate data analysis in Field JG, Clarke KR, Warwick RM 1982, *Mar Ecol Prog Ser* 8: 37-52.

Open Exe nematode abundance, pre-treating the samples with a fourth-root transform (Section 4), and calculating Bray-Curtis resemblances between samples (Section 5). With the latter as the active window, enter the clustering routine, taking **Analyse>Cluster>CLUSTER>**(Cluster mode•Group average) & (✓Plot dendrogram) & (✓Create cophenetic distance matrix), but not the SIMPROF test option for now. (Of course ✓Plot dendrogram would almost always be required).

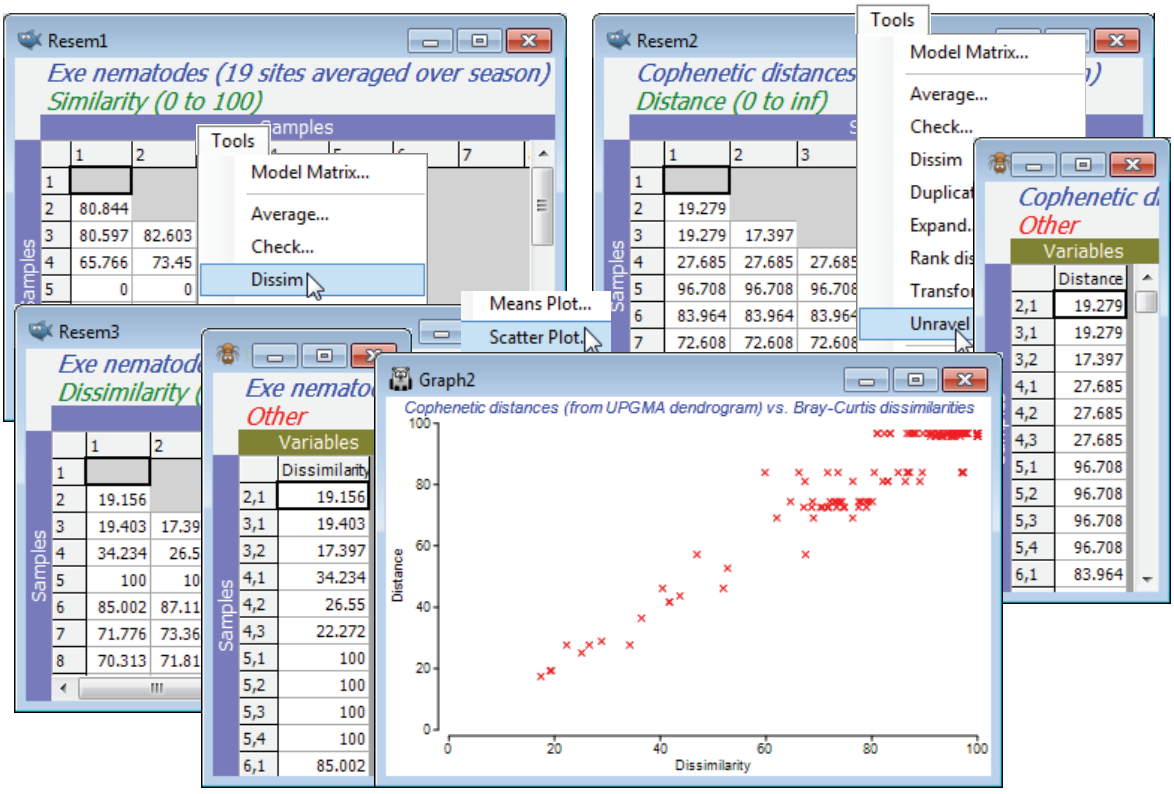
v7



Cophenetic correlation

v7

The dendrogram is displayed in a plot window, and a separate Results window gives a detailed list of the precise similarities at which the groups combine, for this agglomerative method. This also now gives the *cophenetic correlation* (0.958 here), which is a Pearson *matrix correlation* between the entries of the original dissimilarity matrix and the (vertical) distances through a dendrogram between the corresponding pairs of points (the *cophenetic distance* matrix). The closer this Pearson correlation is to 1 the more nearly the dendrogram accurately represents the relationships among the samples in the original (dis)similarity matrix. The concept of matrix correlations is central to several of the PRIMER methods, e.g. in Sections 9, 13, 14 & 17, but is usually computed on the ranks of the two matrices, thus becoming a Spearman matrix correlation. The **Analyse>RELATE** routine in PRIMER 7 now offers Pearson as well as rank-based correlations, so that having taken the option to create the cophenetic distance matrix in the CLUSTER dialog, you could now verify the cophenetic correlation by running RELATE (Section 14) between that distance matrix and the Bray-Curtis dissimilarities. More usefully, you could visualise the relationship by **Tools>Unravel** on both matrices and **Plots>Scatter Plot**, as seen under Unravelling resemblances in Section 5.



Copying & pasting plots externally

Returning to the main point of the previous example, the production of the dendrogram, note that printing or saving dendrograms (and other plots) in a variety of formats is seen in Section 7, but one easy thing to do with any PRIMER plot is to **Edit>Copy** it to the Windows clipboard and then paste it into the slide of presentation graphics software, such as Microsoft Powerpoint. It is then transferred in vector format, so can be 'ungrouped' (e.g. converted to an Office drawing object in Powerpoint) into lines and objects, rather than a pixel-based image. This gives much flexibility for putting the final touches to a plot, e.g. to place titles and plot keys exactly where required, but there is also substantial flexibility in the choice of graphic content within PRIMER itself, as is now seen.

Sample labels & symbols menu/tab

When the active window is a plot, levels of a factor can be displayed in place of sample labels and/or represented by differing symbols, with an accompanying symbol key, using the **Graph> Sample Labels & Symbols** menu. (Alternatively, the same choices result from right-clicking when the cursor is over the plot). If the relevant (✓By factor) check box is ticked, a list of previously-defined factors can be selected from, independently for labels and symbols, so that checking all of (Labels: ✓Plot > ✓By factor) & (Symbols: ✓Plot > ✓By factor) would give a 2-factor annotation of samples on the plot. Note that if the (Labels>By factor) box is not checked, but (Labels:✓Plot) is, then the displayed labels are the sample labels from the resemblance matrix; if the (Symbols>By factor) box is not checked, but (Symbols:✓Plot) is, then a uniform symbol is displayed – this is not relevant for dendrograms but can be useful for other plots. For example, the default symbol for the

above scatter plot has been changed (from a blue square to a red cross) in this way, by clicking on the Symbol: and Colour: icons in the Symbols>Default box of the Samples Labels & Symbols tab of the Graph dialog. The differing shapes, colours etc for the different levels of the chosen factor for symbols can be redefined by the (Symbols:**Key)** button, taking you directly to the Key dialog described in Section 2 under the Factor keys heading. A less direct route is via (Symbols:**Edit**) which takes you back to the Factors dialog (see start of Section 2) which has the same **Key** button. It is also the way of adding new factors from a plot, which are then back-propagated to prior sheets.

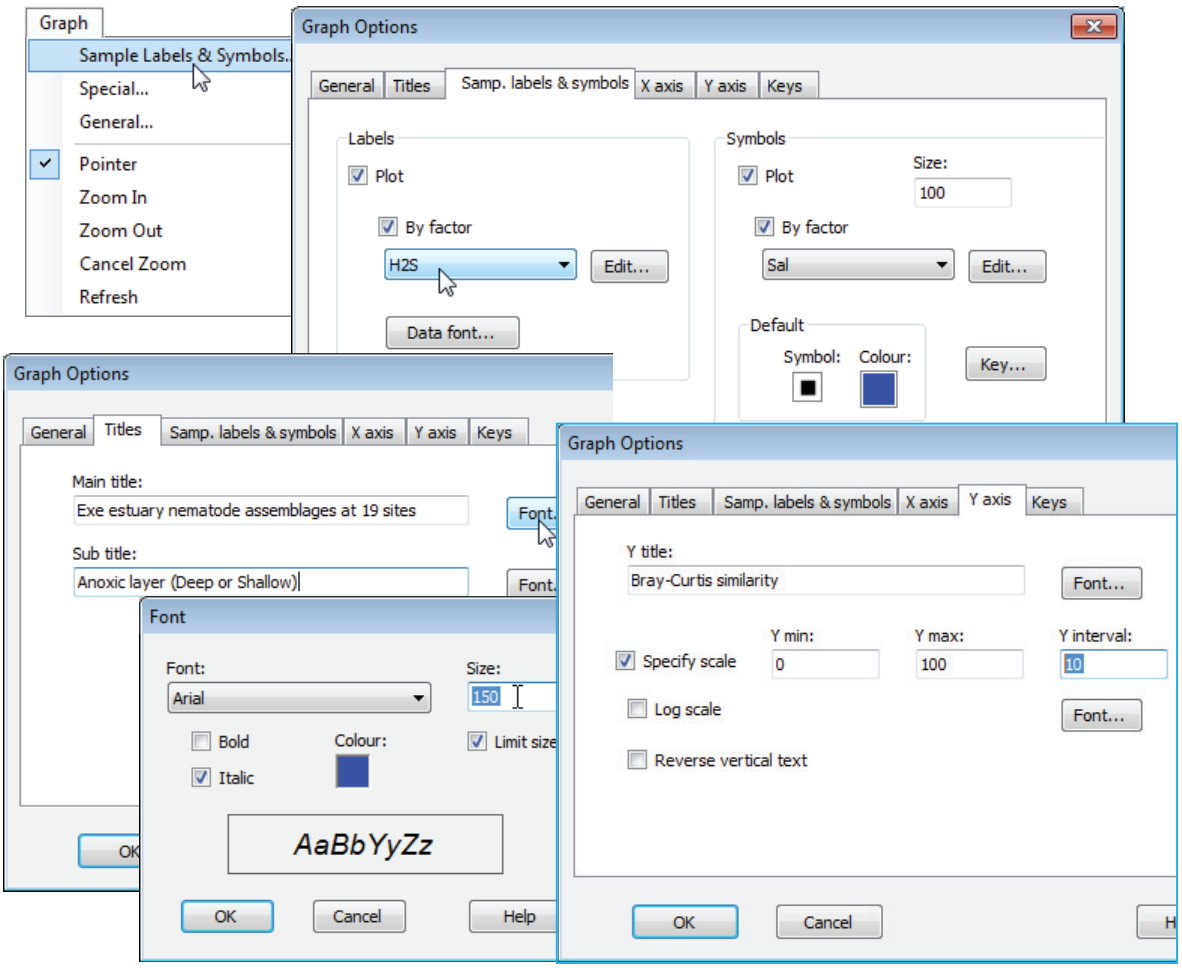
Symbol & text sizes

Label font sizes, typeface, colours etc can be changed with the (Labels:**Data font**) button, and sizes of symbols increased or decreased from the default value of 100 by changing (Size: 100), again in the Symbol area of the **Sample Labels & Symbols** tab – one of the most often used dialog boxes. Note that all such size parameters in PRIMER 7, whether for symbols or text (data labels, main or sub-titles, axes titles and scales etc), are given relative to a default value (usually 100), rather than expressed in terms of a typeface point size, for example. This allows plots to be perfectly scaleable as their windows are resized or printed/saved, without the need for continual redefinition of sizes.

In datasheet Exe nematode abundance, two of the environmental variables from Exe environmental have also been coded as binary factors: interstitial salinity (*Sal*) as Lo (<25%) or Hi (>71% of full seawater); and depth of the blackened anoxic layer (*H2S*) as Shall (<7.5cm) or Deep (≥20cm) – look at these with **Edit>Factors**. As seen in Section 2, there is often a choice of whether environmental variables associated with each assemblage sample are held as a separate data matrix, or as factors within the biological sheet. Here, it is useful to hold some of the data in both forms. With the dendrogram plot (Graph1) as the active window, take **Graph>Sample Labels & Symbols>** (Labels:✓Plot>✓By factor>H2S) & (Symbols:✓Plot>✓By factor> Sal).

Editing plot titles & scales

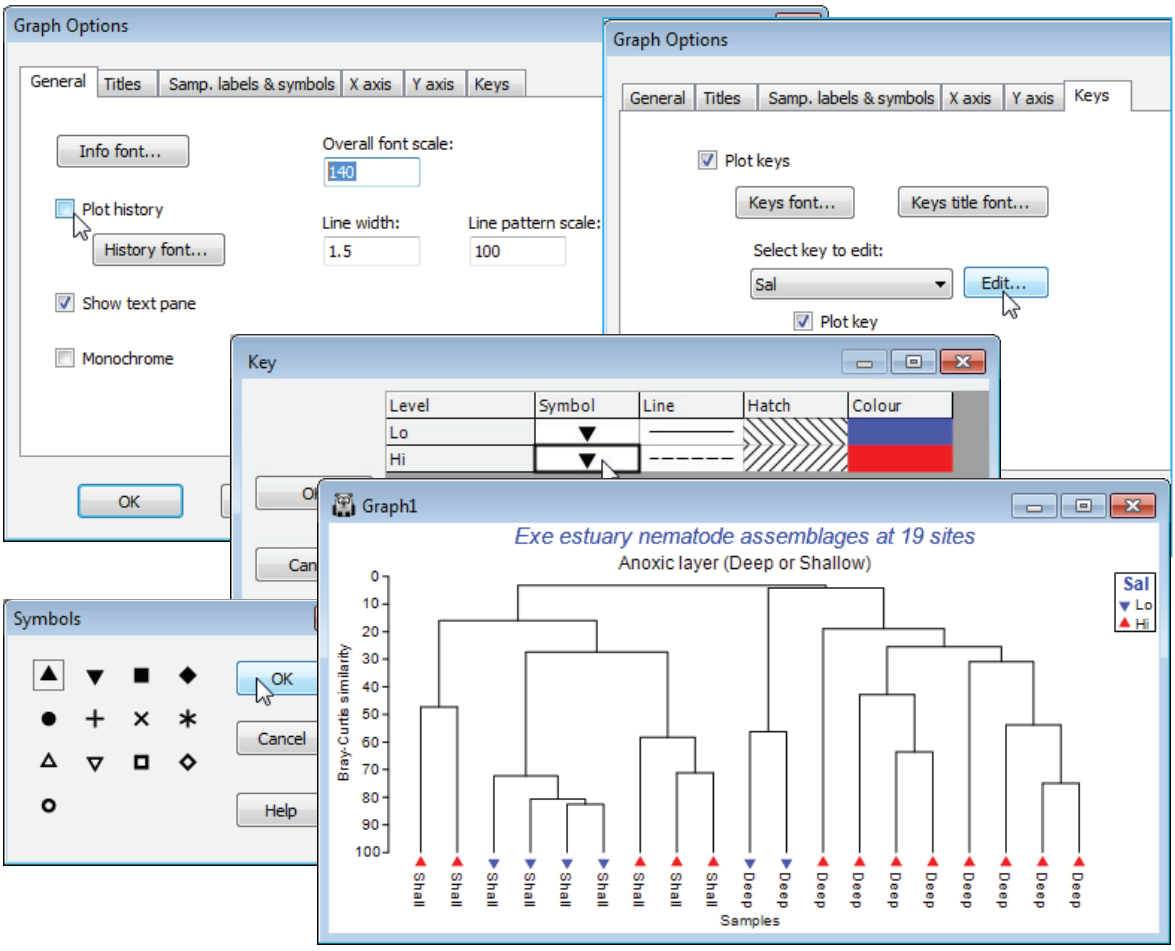
Still in the Graph Options dialog box, take the **Titles** tab and edit the main and sub-title content as shown below, also altering title font sizes and types: (Main title:**Font>**Size:150) & (Sub title:**Font>** Colour:[choose black] & [Italic check box off]). From the **Y axis** tab, change title: (Y title: Bray-Curtis similarity), also respecifying the y axis scale by (✓Specify scale)>(Y interval:10). On the **X axis** tab, take (✓Reverse vertical text).



General menu/tab & Keys tab

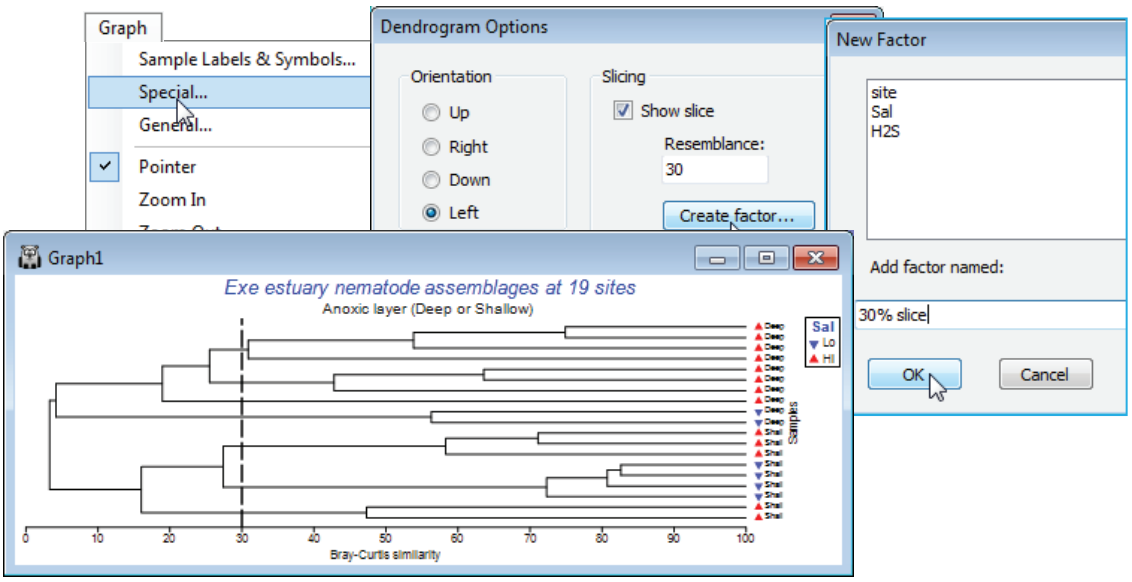
v7

Finally, on the **General** tab (also reached directly from **Graph>General**), thicken up all lines with (Line width: 1.5) , increase the size of all fonts with (Overall font scale: 140), remove the display of the calculation history (transformation, similarity measure etc) by unchecking (Plot history). On the **Keys** tab, take **Keys title font** to make the factor title ✓Bold, and **Edit** to obtain the Key dialog and reverse the upward and downward triangle symbols for this salinity factor (so *Lo* points down!).



Special menu for slicing & orientation of dendrograms

Unlike **Graph>Samples Labels & Symbols** or **Graph>General**, which take you to the Graph Options dialog box, which is displayed in consistent format for these and other appropriate tabs (Titles, X axis, Y axis, Keys), the **Graph>Special** menu item takes you to a specific dialog box applicable only to that type of plot – here a Dendrogram options dialog. This allows selection of orientation, e.g. **Graph>Special>**(Orientation•Left), and a slice drawn through the diagram at a specified resemblance, e.g. by Slicing:(✓Show slice)>(Resemblance:30) & (**Create factor**>Add factor named:30% slice). This creates a factor, levels (a, b, c, ..), of the groups given by that slice.



The newly created factor resulting from the plot will again be back-propagated to any previous data sheet on its direct branch, so whilst it could be utilised to accentuate the clustering structure in this dendrogram, by applying it as the symbols, a more profitable use might be for symbol display on an MDS plot (Section 8), to judge the extent of agreement between clustering and ordination of the same data, under the same resemblance measure. Save the workspace as Exe ws, and close it.

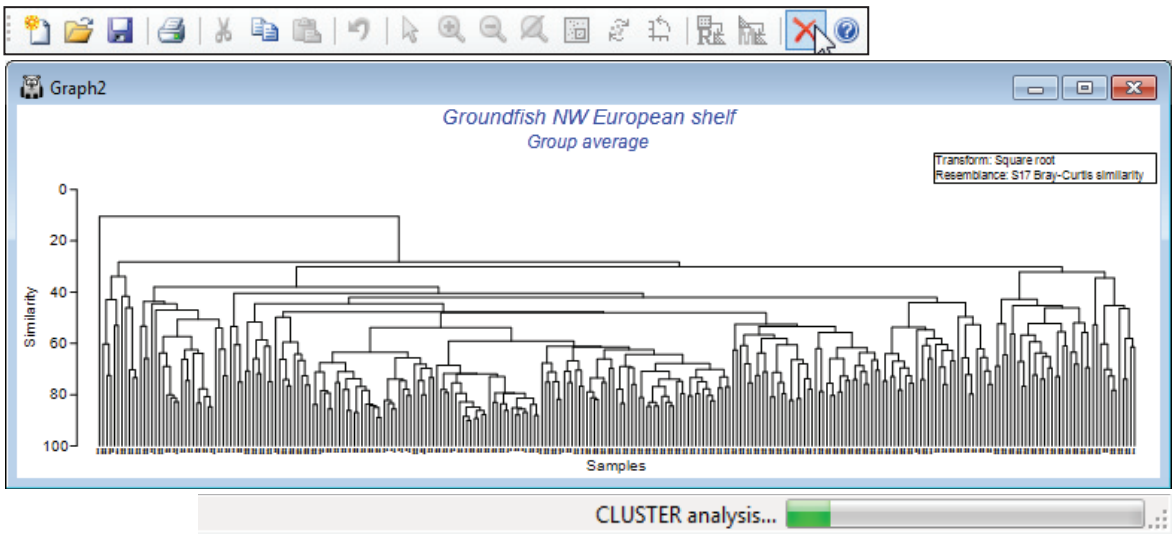
Rotating & condensing dendrograms

The order of samples on the (by default) *x* axis of a dendrogram is to a large extent arbitrary, since all arrangements of samples along the axis, which do not lead to vertical and horizontal lines intersecting, are equally satisfactory displays – think of the dendrogram as a 'mobile', of horizontal rods and vertical strings, which can be rotated at will. Such rotations can be achieved by clicking on any of the horizontal 'rods' and, whilst it is not appropriate to use this feature to re-arrange the samples close to some desired *a priori* sequence(!), it can be useful in displaying visual agreement between clusters from different analysis choices, or comparing abiotic and biotic groupings for the same set of samples. Clicking on vertical 'strings' collapses the clustering under the selected point, replacing it with a single dashed (green) line, to indicate the presence of condensed structure. These lines are labelled with capital letters within a *text pane*, below the plot, which defines the samples contained in a hidden structure (suppressing the text pane is possible, by **Graph>General**). For dendrograms with many samples, this feature should make it possible to view the overall (coarse-level) structure, and the fine-level grouping can then be seen by zooming in on areas of the original dendrogram.

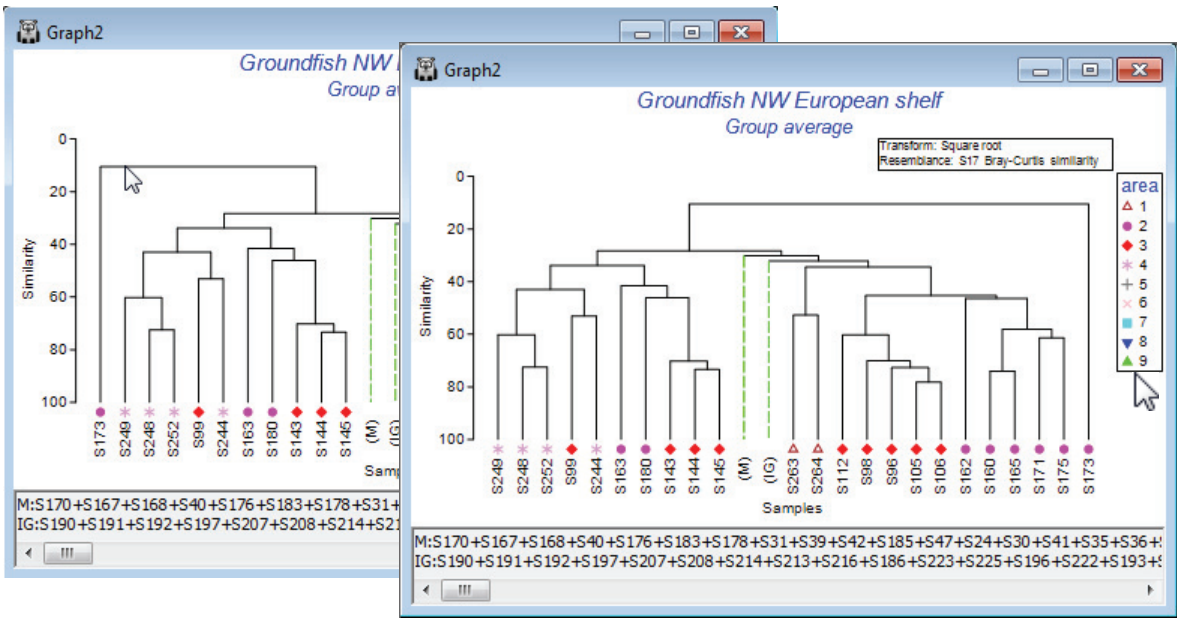
Re-open the workspace Groundfish ws from C:\Examples v7\Europe groundfish, met in Section 5, with datasheet Groundfish density of 277 samples of 93 groundfish species, captured in research trawl surveys of 9 areas of European shelf waters (factor *area*). Produce a dendrogram based on Bray-Curtis similarities from square root transformed densities, with **Pre-treatment>Transform (overall)>**(Transformation:Square root), **Analyse>Resemblance>**(Analyse between• Samples) & (Measure•Bray-Curtis similarity), **Analyse>Cluster>CLUSTER>**(Cluster mode•Group average). Alternatively, take (Cluster mode•Single linkage) for a clear demonstration of why group average linkage is generally superior to the 'chaining' that single linkage produces (see CiMC Chapter 3).

Timing bar, Stop Tasks & multi-tasking

As discussed at the start of this section, if you have set SIMPROF running, with the (✓SIMPROF test) check box, you will find that calculating the dendrogram takes some while – the timing bar on the Status Bar at the bottom of the PRIMER desktop (turning green, as the calculation progresses) scarcely seems to move. An example of this size (277 samples) will complete in a not unreasonable time but if you embark on a calculation which is clearly going nowhere, execution can be stopped cleanly, without damaging the workspace in any way, by clicking on the Stop Tasks icon  , on the Tool Bar (equivalently, take **Tools>Stop Tasks**). The PRIMER environment is also fully multi-tasking – if a calculation is set to take a long time, you can run other, less intensive, manipulations simultaneously within the same workspace, with no fear that they will interact with each other. However, a Stop Tasks instruction will terminate all routines currently running in parallel in that workspace. Of course, multiple runs of PRIMER 7 can also be launched and will operate quite independently of each other – there is no live linkage to files or workspaces external to the current workspace (files can only be transferred between workspaces by **Save** from one workspace and **Open** in the other, and it is always a copy of the file contents which is taken into the workspace).



Using **Graph>Sample Labels & Symbols**, add symbols for factor *area*, but the large number of samples makes the symbols too small to see. So, click on (say) two of the vertical lines to collapse the large middle section of the tree. Note the appearance of a text pane, listing the samples in the condensed branches. Also see how the structure may be arbitrarily rotated by clicking on, say, the top horizontal line, to rotate the sample S173 to the other side of the tree.



Ordering factor levels in keys

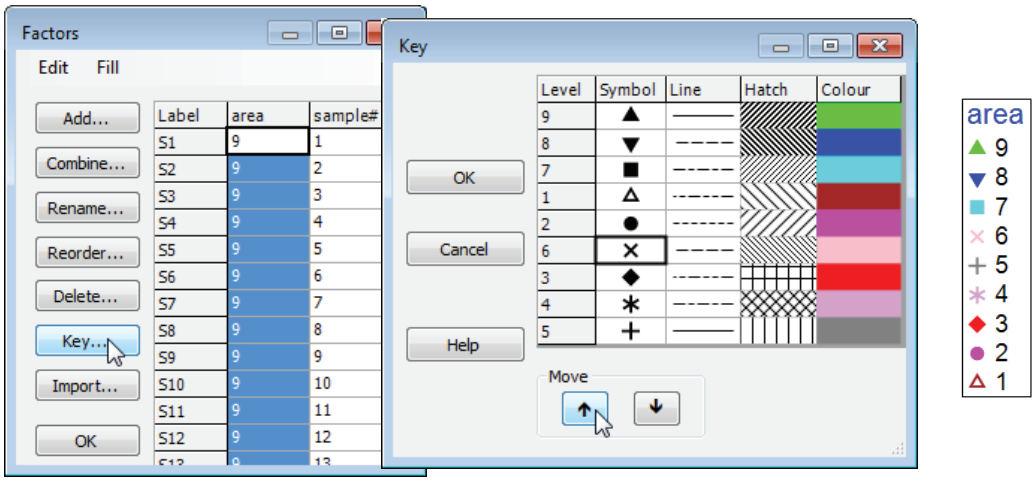
v7

PRIMER 7 now automatically displays the levels of a numeric factor in increasing order in a plot key, but note that it makes no attempt to order non-numeric levels alphabetically, instead keying them in the order in which they are met in the factor sheet, which is very often the order in which they should naturally be presented (think Spring, Summer, Autumn, Winter!). The key ordering can be manually overwritten in either case. Here, if it was natural to present areas in the reverse order (you will see from the factors sheet that *area* 9 samples, Bristol Channel, are the first in the matrix) then go to the Key dialog (by the **Key** button in the Factors dialog or on **Graph>Sample Labels & Symbols**), and a set of (Move>↓) & (Move>↑) operations re-arranges levels in any desired order.

Point & click short-cuts

v7

There are often several ways of getting to the same dialog in PRIMER and the third, and quickest, way to bring up the Key dialog is simply to click on the key itself, as shown in the plot above. This is a generic new feature in PRIMER 7: click on any peripheral structure of a plot (Key, Titles, *X* or *Y* axes, History box) and the appropriate dialog box will appear.



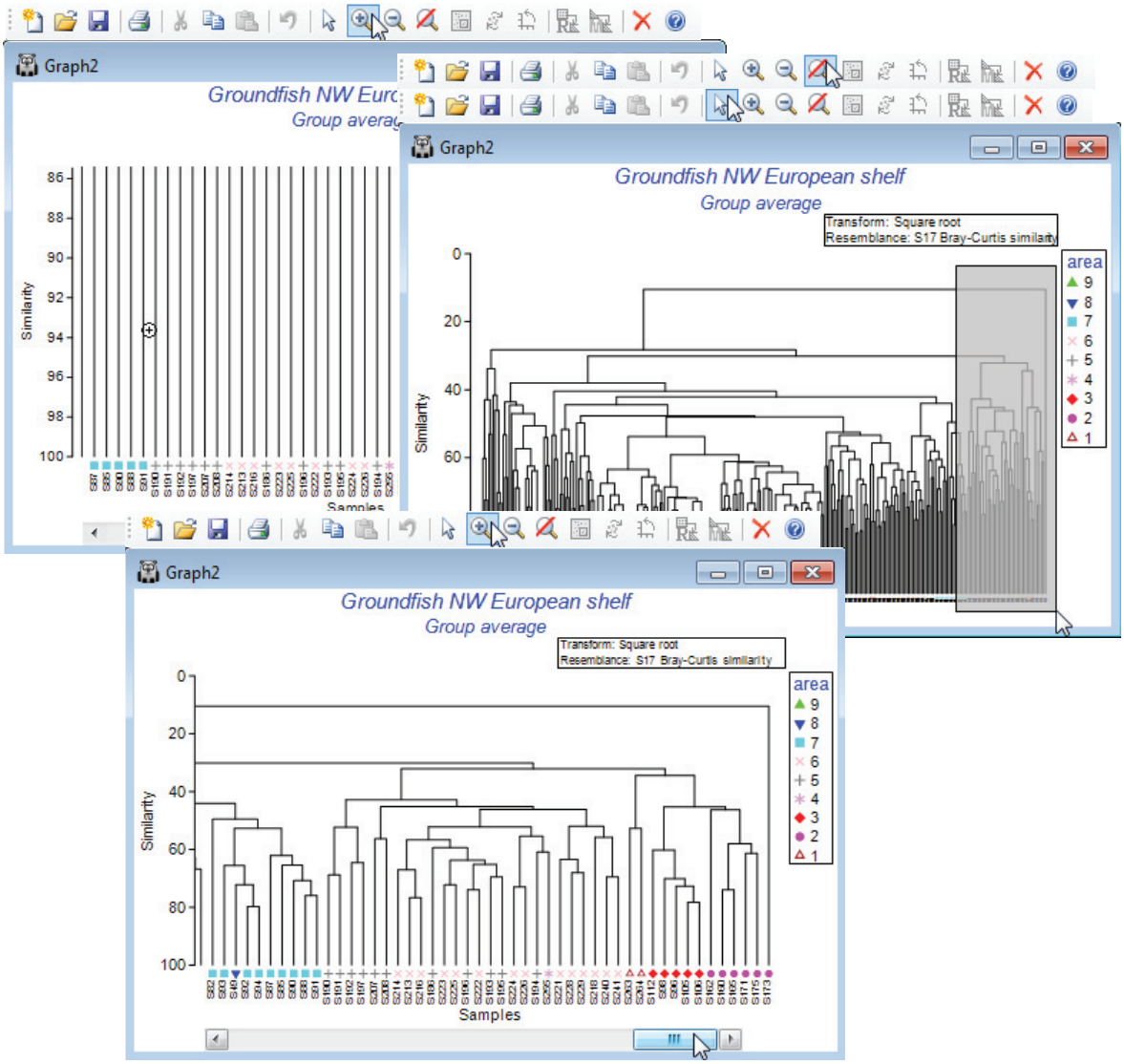
Zooming dendrograms

Zooming is invoked by **Graph>Zoom In** or **Zoom Out** from the main menu, or by clicking on the Zoom in or out icons on the Tool Bar. The cursor changes to ⊕ or ⊖ when over the plot, and left-clicking zooms one step in or out. To leave the plot in its current (possibly zoomed) state and return to default operation, click on the pointer icon on the Tool Bar, or take **Graph> Pointer** (remember that the graph menu can also be obtained at any time by right-clicking when the

v7 cursor is over the plot). To restore the plot to its original, unmagnified state, click on the new cancel zoom icon on the Tool Bar, or select **Graph>Cancel Zoom**.

Instead of zooming by incremental steps, you can go straight to the final zoomed area by drawing a box around the area to magnify: with the cursor in the usual pointer mode (click on on the Tool Bar if necessary), draw a box by left-clicking and holding at one corner and dragging over the required rectangle, then releasing, in usual Windows fashion. A single click on the icon on the *v7* Tool Bar (or **Graph>Zoom In**) will take you straight to the zoomed area. (The process is reversed, as above, by taking **Cancel Zoom** or its icon). But note that, unlike the incremental zoom, which preserves the *aspect ratio* (the displayed *y*:*x* axis ratio) of the diagram, a rectangular zoom will change the aspect ratio so that all the information within the box is magnified into the current size of window, however long and thin (or short and fat) the drawn rectangle originally was. This is a powerful feature for zooming on dendrograms, since a long, thin rectangle allows you to view a small subset of the samples (*x* axis) across the whole similarity scale (*y* axis). Under zooming, note that the axes are always shown, even when the zoomed area is well away from them, and scroll bars are displayed on the axes. By dragging these scroll bars back and forth (or up and down) the whole tree can be viewed, piecemeal, at the current aspect ratio and magnification.

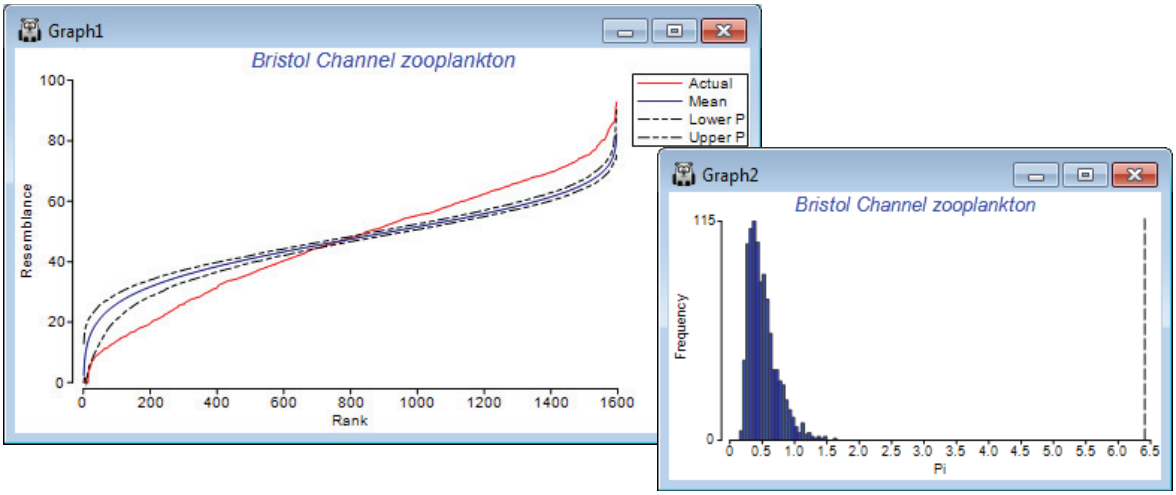
Reverse the condensing of the middle section to reinstate the full tree (rotating and collapsing are *toggles*, switched on and off by repeated clicking on the same line) and now try to zoom in on the fine detail. Repeated use of the ⊕ cursor from the Tool Bar icon is not effective. By the time the symbols are visible, the similarity scale is too narrow to see the clustering structure. What is needed is a change in aspect ratio: cancel the zoom, change to the pointer , draw a tall narrow box over part of the dendrogram, and **Zoom In** again ( ). A viewable dendrogram now results, which can be scrolled across, using the horizontal scroll bar. Save the Groundfish ws workspace and close it.



SIMPROF method

The similarity profile test (SIMPROF), Clarke KR, Somerfield PJ, Gorley RN 2008, *J Exp Mar Biol Ecol* 366: 56-69, is a permutation test of the null hypothesis that a specified set of samples, which are not *a priori* divided into groups, contain no multivariate structure to further examine. (Do not confuse this with the ANOSIM test, Section 9, which tests prior group structures of times, sites, treatments etc). The *SIMPROF procedure*, usually a sequence of SIMPROF tests, is used extensively in PRIMER to provide stopping rules for all the clustering methods: unconstrained sample clustering in this section (and Chapter 3, CiMC p3-6); species (or more general variable) clustering into *coherent* response curves in Section 10 (and the start of Chapter 7, CiMC); and biotic sample clustering constrained by thresholds on environmental variables in Section 13 (and Chapter 11, CiMC p11-13). The *similarity profile* itself is the set of resemblances among all pairs of the specified samples, ranked from smallest to largest, and the ordered resemblances then plotted (*y*-axis) against their rank (*x*-axis). The departure of this curve from its 'expected' shape under the null hypothesis is the basis of the test. For example, if there is genuine clustering within a set of biotic samples, there will be many more smaller similarities and larger similarities than if all the samples came from the same community (and therefore all had intermediate similarities to each other). The 'expected' profile is obtained by permuting the entries for each variable (e.g. species) across that subset of samples, separately for each variable, thus producing a 'null' condition in which samples can have no real structure. Such simulations realistically fix the variable values, e.g. to have the same pattern of rare and common species, with the same counts, as the real matrix, and thus require no assumptions about the differing forms the distributions of abundances may take for the differing species. The random rearrangements are repeated a large number of times (under user control), producing many 'expected' profiles under the null, for which the average and percentile (say 95% or 99%) values at each rank are plotted along with the real profile. A typical real profile, with mean and 99% limits from the permuted profiles, for all 57 samples of the data below, now follows (on left; see later for the routine which constructs these plots, under SIMPROF direct run).

v7



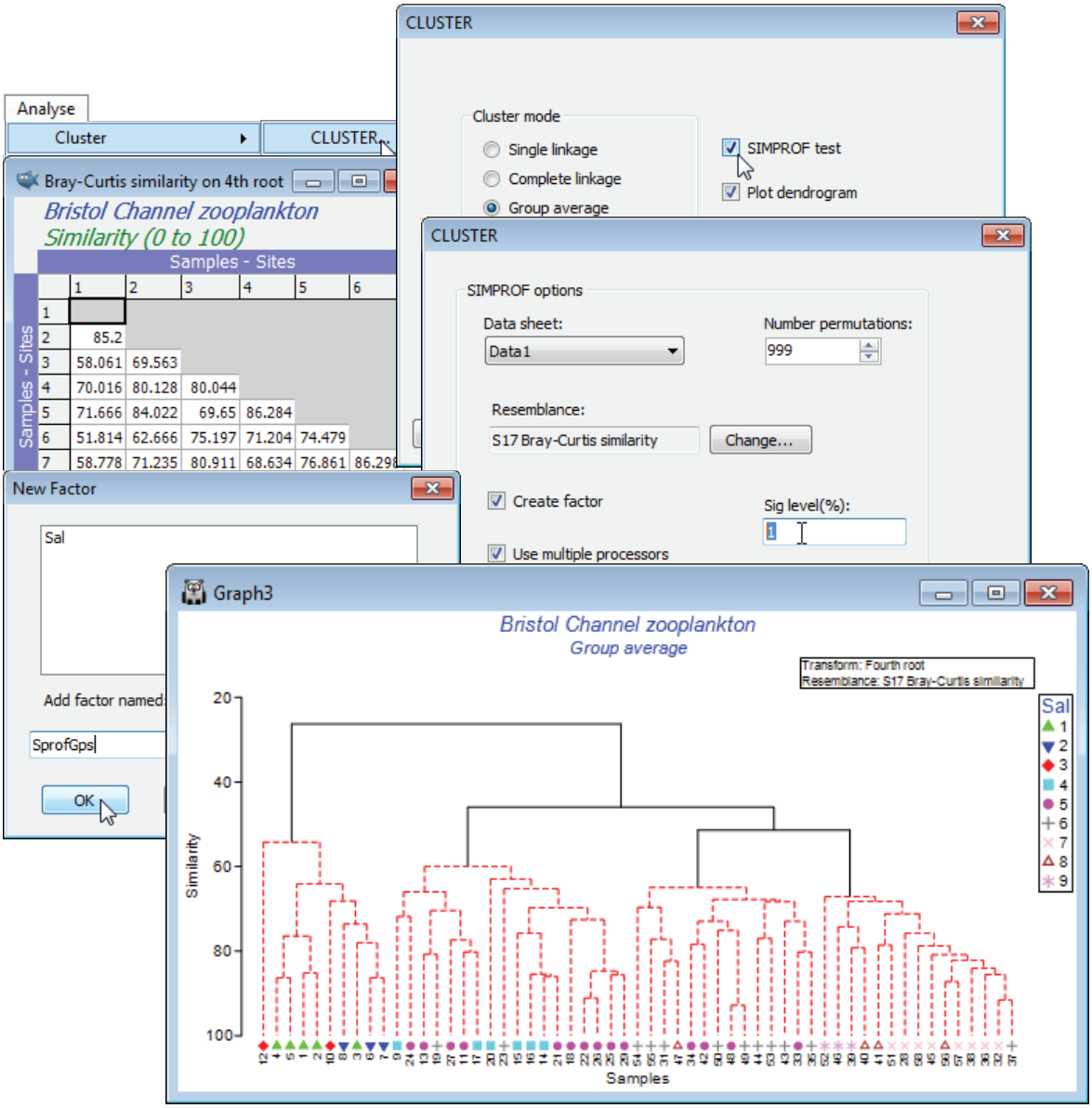
The summed absolute distances (π) between the real similarity profile and the simulated mean profile is the test statistic. A second set of simulated profiles are then generated and π computed between each of these and the mean profile (from the first set). This defines a range of likely values of the test statistic π under the null hypothesis (above histogram, right), and the real π (dashed line, far right) is compared to this to give a *p* value, as for any test, given as a percentage (see stages in permutation testing, Chapter 6, CiMC). Here the real π is the most extreme of 1000 arrangements of the matrix (999 permuted and one real one), hence $p < 1$ in 1000 (0.1%) and the null is rejected – there is structure. The SIMPROF procedure in CLUSTER separately repeats this test on the two sample clusters at the next level down, and so on until no further significant results are obtained.

(Bristol Channel zooplankton)

Densities from 24 species of zooplankton at 57 sites in the Bristol Channel and Severn Estuary, collected by double-oblique net hauls, are in C:\Examples v7\BC zooplankton\BC zooplankton density(.pri). The sampling sites were defined as a grid (Fig 3.2, CiMC), and samples taken through time over a single year and averaged to give one seasonally-averaged sample per site. There is thus no prior structure of groups and replicates within groups (though there is a natural salinity gradient, described by factor *Sal*, with 9 numeric levels). The original data is from Collins NR & Williams R 1982, *Mar Ecol Prog Ser* 9: 1-11, who identify four main clusters of sites.

It is relevant to ask what the statistical evidence is for there being such a division at all, and if so, how much of the group structure can justifiably be interpreted. Open the data file BC zooplankton density and generate the cluster dendrogram from Bray-Curtis similarities on 4th root transformed densities (as in Sections 4 & 5), then **Analyse>CLUSTER>**(Cluster mode•Group average), but this time taking the option (✓SIMPROF test). Look at the dialog under the **SIMPROF** tab, though the defaults probably be taken for (nearly) all: the matrix whose species rows will be independently permuted is Data1, the 4th root transformed data; no other choice than Resemblance: S17 Bray-Curtis similarity makes sense on the randomly permuted matrices since that was the choice on the real matrix; the % significance level is conventionally taken as 5 though could be more stringent, given ultimately that 7 tests are performed here, so change it to, say, 1; the 999 permutations will typically be sufficient (for computing the mean, and a further 999 for departures π from the mean) bearing in mind the computation needed to recompute and re-order the $n(n\text{-}1)/2$ similarities (with n samples) for each permutation, and then repeat this through the dendrogram; and clearly the use of multi-cores in the processor is beneficial to that. The final group structure, from the series of SIMPROF tests, is placed in a factor, generating another dialog, (Add factor named: SprofGps).

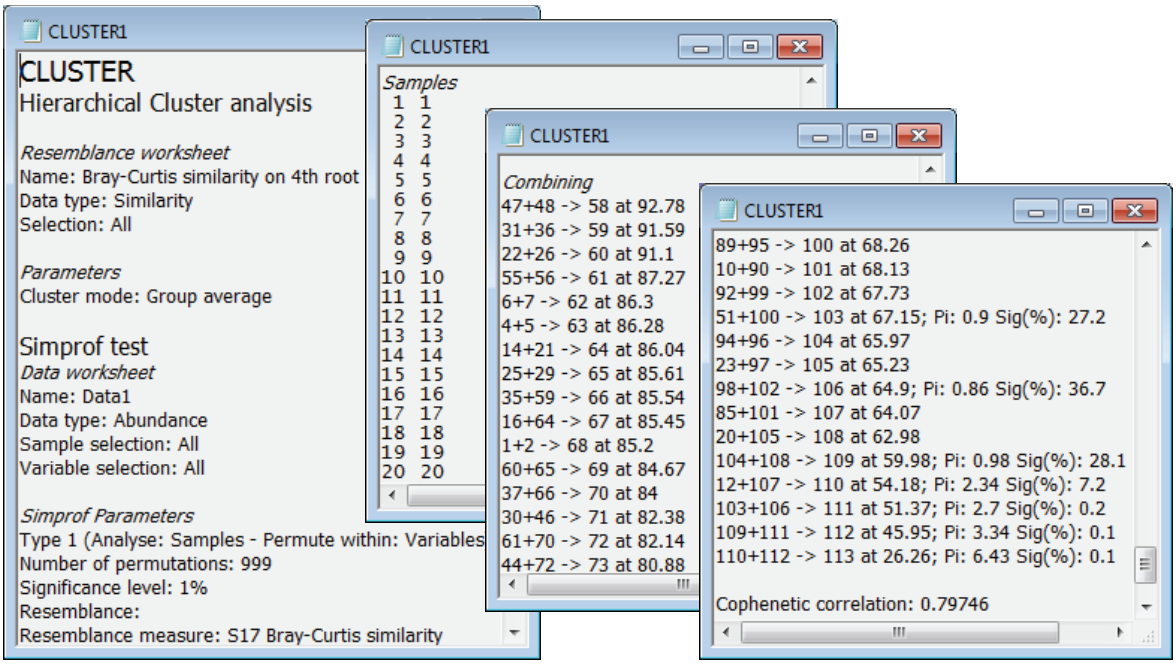
v7



With n only 57 in this case, and with few tests needed, the SIMPROF procedure runs very quickly. The dendrogram shows the four groups of sites identified by Collins & Williams but now with firm statistical support: the black lines indicate groups that are established, with red lines showing a sub-structure from the clustering for which there is no statistical support from SIMPROF to permit interpretation. That the groups bear a strong relation to salinity is seen by displaying the salinity factor as a symbol, with **Graph>Sample Labels & Symbols>**Symbols:(✓Plot)>(✓By factor:Sal).

CLUSTER results window

In addition to the dendrogram plot itself, **Analyse>CLUSTER** (like all analysis routines) produces a separate Results window (here CLUSTER1) which firstly lists the conditions under which the analysis was run (e.g. whether on a selection of the matrix, with what linkage option etc), and then outputs text-format information. For succinctness, the Results windows will often use the sample numbers (1-57) rather than the sample labels (stations 1-29, 31-58, confusingly, since station 30 was not sampled!), so a listing is initially given of the numbers and their corresponding labels (the last label here, of sample 57, thus being station 58). Then the results specify, numerically, how the dendrogram is constructed, just in case the precise numbers are needed for another purpose: sample numbers 47 & 48 (stations 48 & 49) are the first to group, at similarity 92.78, with the new group labelled 58, then 31 & 36 group at 91.59, …, 16 & 64 (i.e. 16 & 14 & 21) at 85.45 etc. Likely to be most useful here, however, are the SIMPROF test results. These are read from the bottom upwards: π =6.4 (p<0.1%, its most extreme value for 999 permutations) for a test that all samples are from the same assemblage; and π =3.3 & 2.7 (p<0.1% or 0.2%) for the successive splits, at 46.0% and 51.4% similarity, of the three right-hand groups. Site 12 is borderline for splitting from the rest of the left-hand group, at 54.2% similarity (π =2.3, p<7%), but there is no evidence for the apparent division of the second group into two at 60.0% similarity (π =1.0, p<28%), or any of the other groups. Tests of finer-level structure are not carried out, if the differentiation of the coarser level structure is not significant, so only seven tests are needed here. Note that the choice of threshold significance level (p<1%) for rejecting the null hypothesis of 'no structure' is not at all critical here – p<5% or p<0.5% would have led to the same set of decisions – and such robustness is common.



SIMPROF direct run

v7

SIMPROF can be run directly using **Analyse>SIMPROF**, rather than as part of another analysis such as CLUSTER (above), UNCTREE or kRCLUSTER (later this section) or LINKTREE (see Section 13). In that case, the active window must be the data sheet, the rectangular matrix whose variables are permuted randomly and independently across the samples. SIMPROF must always have such an underlying data matrix available – it cannot work solely on a triangular resemblance sheet. Thus when the SIMPROF option is taken in CLUSTER – which is run when the active window is a triangular matrix – PRIMER uses its internal knowledge of how that resemblance matrix was calculated to specify the correct data matrix, as a default for (Data sheet:) under SIMPROF options. Change this default at your peril! – its main purpose is simply to remind you that SIMPROF always works on the underlying rectangular array not the triangular matrix.

Direct runs of SIMPROF are used to test for evidence of internal group structure in the full set of samples that are submitted to it, i.e. a single test rather than the (usually large) series of subset tests in the CLUSTER option. The advantage of doing a single test at a time is that more information can be output, as seen in the plot windows shown above under the SIMPROF method heading, for a preliminary test of any structure in the full set of 57 samples for the Bristol Channel zooplankton.

Another output option for **Analyse>SIMPROF**, selected by checking ✓Stats to worksheet, is of the data used to plot the similarity profile itself. This worksheet will have a number of rows equal to the number of entries in the resemblance matrix, containing as 'variables': the real ranked similarities; the mean similarities from the permutations; the lowest and the highest similarities obtained, at each rank, over all permutations (not shown on the plot); and the lower and upper 99% limits (or whatever % specified) of the permuted values at that rank.

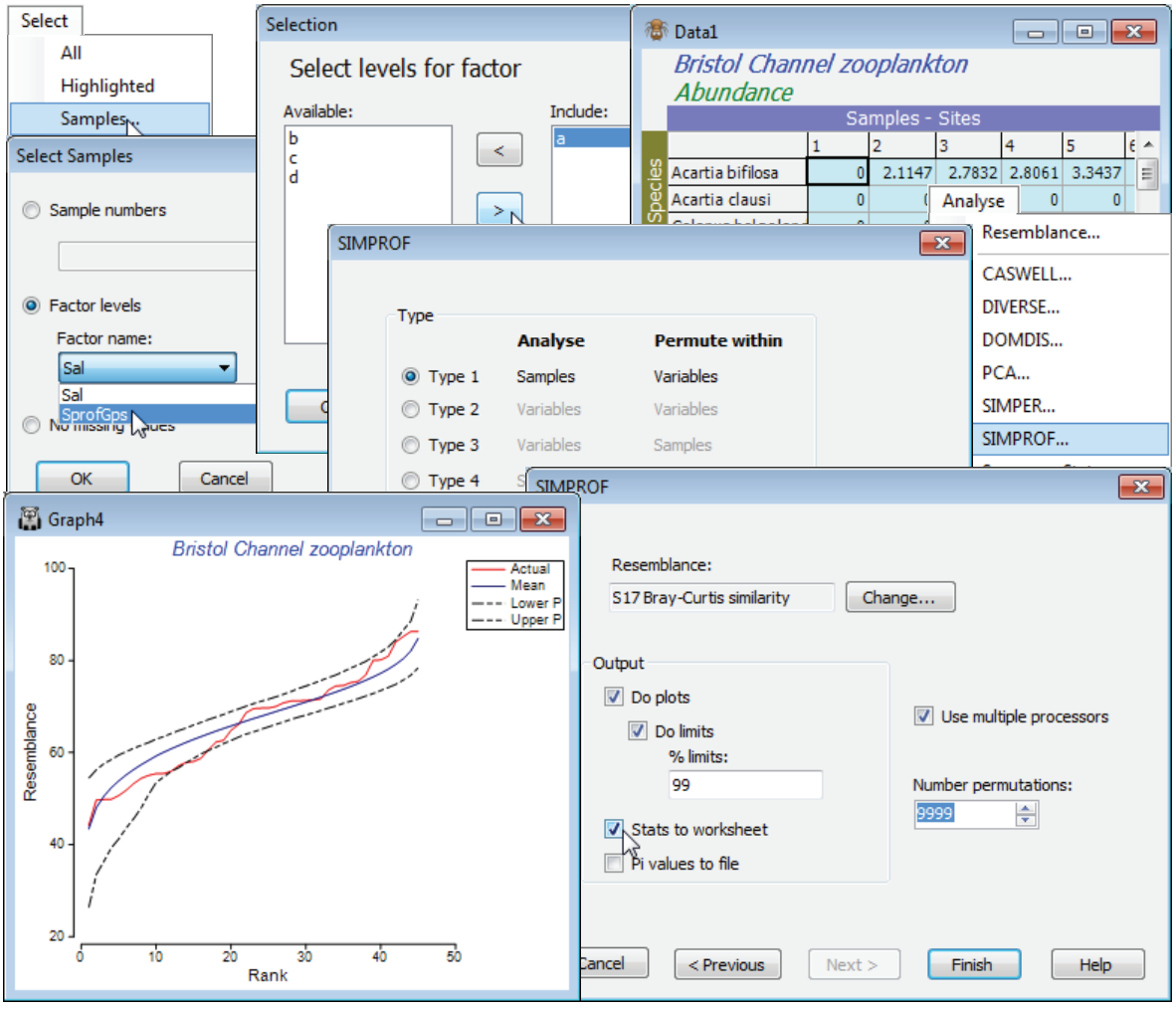
SIMPROF Types (1-4)

v7

The first dialog box from running **Analyse>SIMPROF**, however, is of a new option to PRIMER 7, a choice of 4 types of SIMPROF test, which cover all 2 × 2 combinations of analysing samples or variables and permuting within samples or variables. The default, described above, is now referred to as a *Type 1* test, in which similarities are calculated between all pairs of samples and the profiles recalculated under the null hypothesis by permuting entries of the data matrix separately within variables, across the samples. *Type 2* and *Type 3* SIMPROF tests concern analysis of variables in which, for example, the profile consists of index of association values calculated among species. Dependent on the hypothesis being tested this can either involve again permuting within variables across samples (Type 2) or within samples across suitably standardised species (Type 3). Chapter 7 of CiMC gives the motivation and examples for both Type 2 and 3 tests, seen again in Section 10, based on Somerfield PJ & Clarke KR 2013, *J Exp Mar Biol Ecol* 449: 261-273. The final option (*Type 4*) in this dialog, analysing samples and permuting within samples, has been included purely for completeness but has not been described and seems less likely to be practically useful.

SIMPROF on a subset of samples

From the Bristol Channel zooplankton transformed data matrix Data1, select (say) the samples of the first of the four groups (*a*) produced by the above series of SIMPROF tests under CLUSTER, by **Select>Samples**>(•Factor levels)>(Factor name: SprofGps)>**Levels**>(Include: a), and with this sheet as the active window, take **Analyse>SIMPROF**>(Type•Type 1). Take defaults on the next dialog except (Number permutations: 9999) & (✓Stats to worksheet). As already seen, $\pi = 2.3$ and $p = 7\%$ for this test but the direct run of SIMPROF gives further graphical information, below.

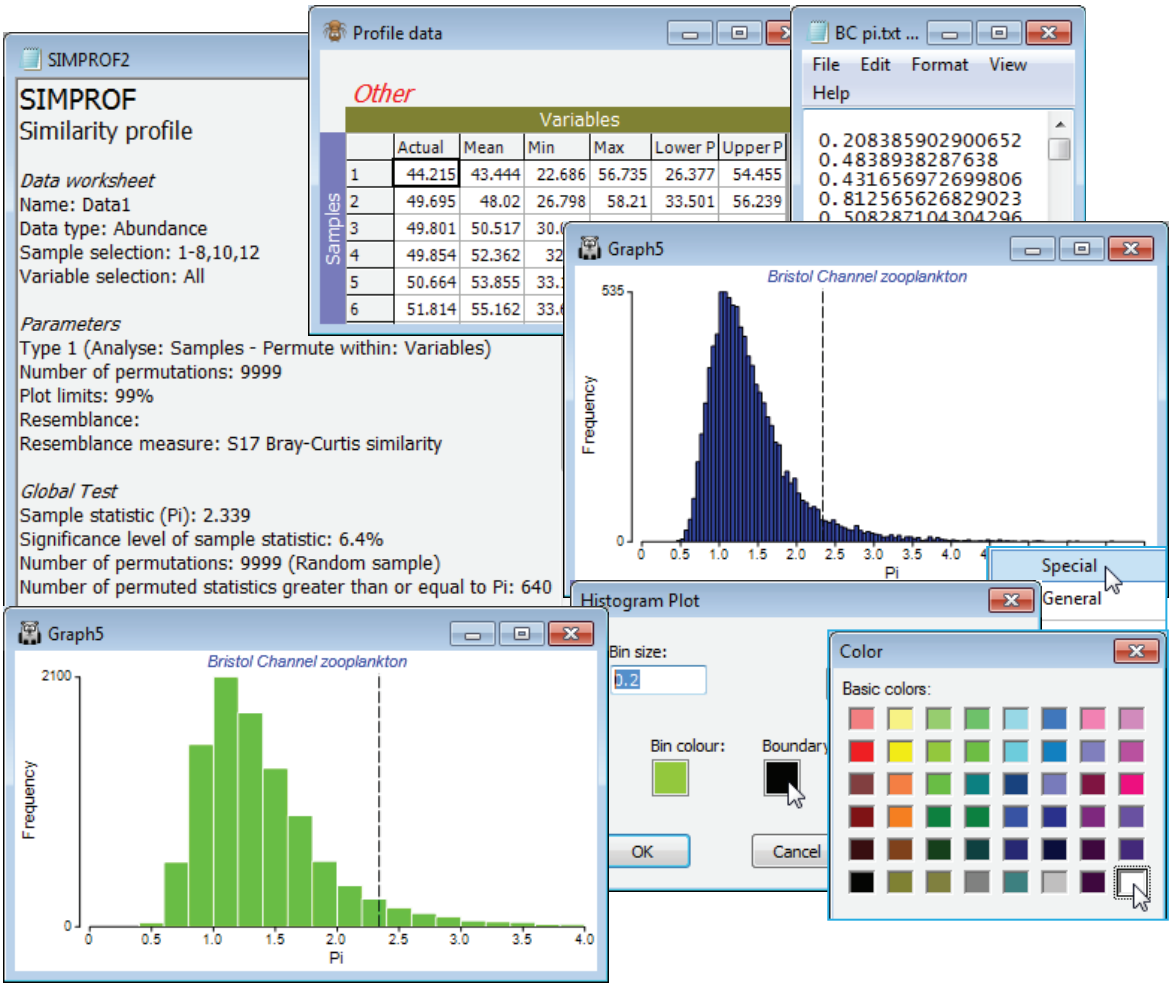


The previous SIMPROF plot, for the full set of 57 samples, showed an excess of both smaller and larger similarities in the real profile than would be expected by chance, if there were no structure in these data. In contrast, the real similarity profile for this subset of sites 1-8, 10, 12 lies almost fully within the envelope of the 99% limits for the permuted similarity profiles at each rank. Of itself, this juxtaposition of curves is not the test of departure from the null hypothesis; since there are 45 similarity ranks from 10 samples, there is a fair probability that a 1 in 100 event (the probability that a point lies outside its 99% limits by chance) occurs at least once in 45 'trials'. Hence the use of a test statistic which is the summed absolute distances, π, between individual profiles and the mean of the profiles under permutation. If the real profile is further from that mean than 95% (say) of the individual permuted profiles are, then this is evidence to reject the hypothesis of no structure.

Histograms of null distributions

As in all permutation tests in PRIMER v7 (e.g. in ANOSIM, RELATE, BEST etc), a further output from **Analyse>SIMPROF** is thus a histogram of the values of the test statistic (π, here) for the null hypothesis conditions, under permutation, with the real value ($\pi = 2.3$) also indicated. The relevant p value is given in the SIMPROF results window – not 'significant' but borderline here, $p \approx 6.5\%$. Minor variations of both π and p values, from those in the results window for the earlier CLUSTER run, are to be expected because the permutations are random – each new run gives slightly different answers. Previously, the default of 999 permutations was selected. If greater precision is needed in significance levels, then the number of permutations should be increased, as here to 9999 (there are a vast number of possible permutations for this particular test). Binomial calculations show that tests with a true p of 5% will return a p in the range (3.5%, 6.5%) with 999 permutations, and in (4.6%, 5.4%) for 9999 permutations – this is true for all tests employing random permutations.

v7 As previously seen (Section 4) with direct plotting of histograms, via **Plots>Histogram Plot**, the **Special** menu (right click when over the plot) allows changes to bin width, colour, axis scales etc. A final option on the SIMPROF dialog (✓Pi values to file) is to export the π values from all the permutations to a text file (*.txt). This is not commonly used but is again an option with most permutation tests in PRIMER; e.g. it would allow the user to redraw the null distribution histogram of π with other software, or perhaps examine its parametric form etc.



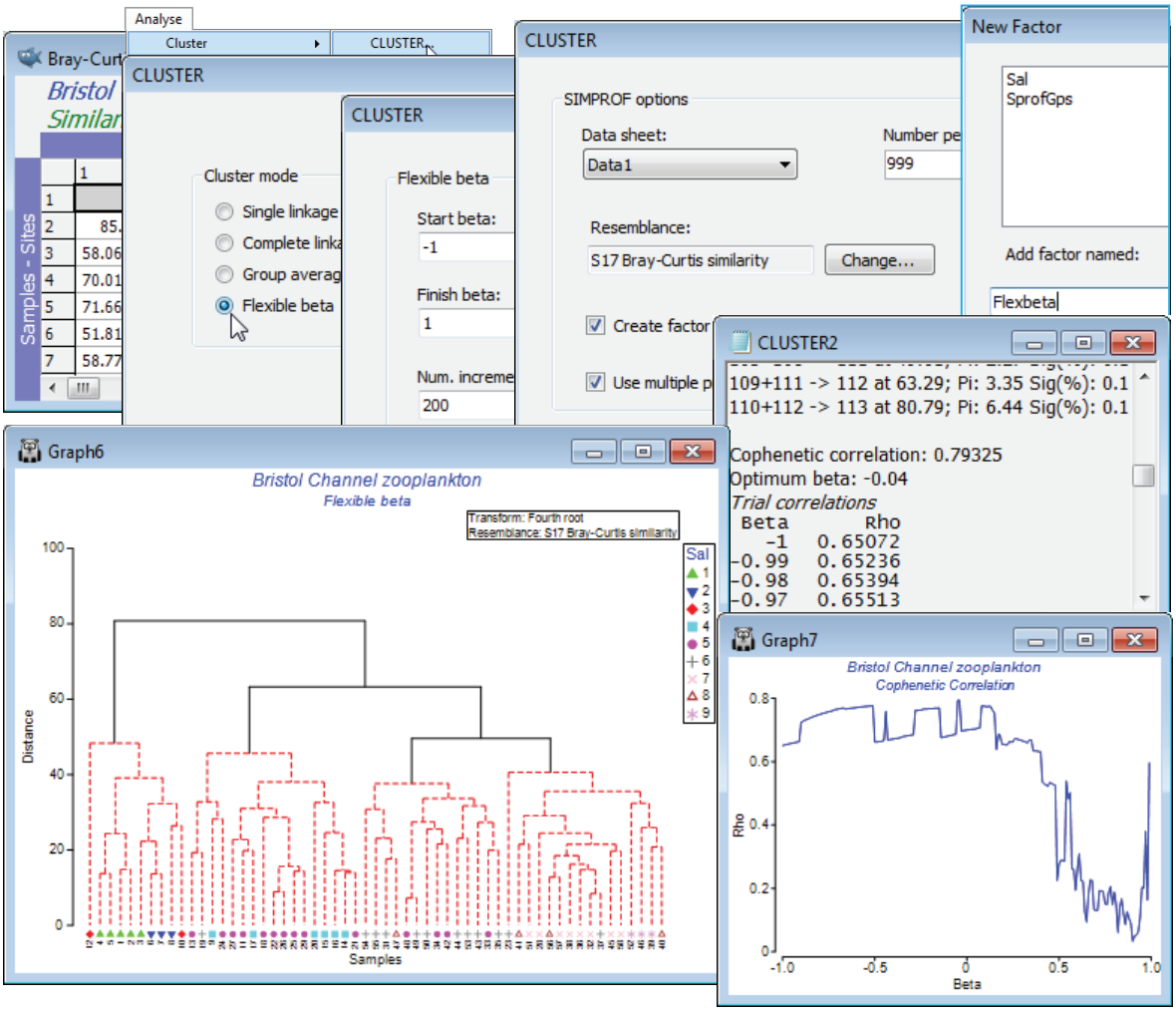
Linkage by flexible beta method

There are four possible Cluster mode choices within the **Analyse>Cluster>CLUSTER** dialog box, distinguished by the way they redefine the among-group dissimilarities at each proposed step of the agglomerative process. The *linkage options* are: •Single (/nearest neighbour) linkage, which has a tendency to produce unhelpful 'chaining' of groups, with many steps adding just a single sample to an existing group; •Complete (/furthest neighbour) linkage, which tends to have the opposite 'over-grouped' effect; •Group average (Unweighted Pair Group Method with Arithmetic mean UPGMA) which is the option shown in all the above plots and is widely used; and •Flexible beta, introduced by Lance GN & Williams WT 1967, *Comp J* 9: 373-380, a generalisation of a WPGMA method in which a range of options is controlled by choice of a parameter β. Chapter 3 of CiMC gives precise definitions of all these options, e.g. for flexible beta see the footnote on p3-4. Choice of β is made automatically to maximise the *cophenetic correlation* ρ (rho) between the dissimilarities/distances in the resemblance matrix and distances through the dendrogram between the matching pairs of samples – this idea was met near the beginning of this section – and a plot of ρ *vs.* β displayed.

v7

Remove the selection on the fourth-root transformed data matrix Data1, by **Select>All** (and **Edit> Clear Highlight**, though this is not essential) then with the active sheet as the similarity matrix calculated from Data1, take **Analyse>Cluster>CLUSTER**>(✓SIMPROF test) & (Cluster mode• Flexible beta)>(Start beta: –1) & (Finish beta: 1) & (Num. increments: 200). These are the defaults, meaning that the cophenetic correlation is computed and graphed for β in increments of 0.01, with the optimum β (maximum ρ) given in the Cluster results window, and this value used to calculate the dendrogram. Note that β does need to be in the range (–1, 1) but negative values (or zero) make better sense theoretically, as is seen here in the line plot of the cophenetic correlation ρ *vs.* β, so there is a case for restricting to (Start beta: –1)&(Finish beta: 0)&(Num. increments: 100). If a fixed value of β is preferred (Lance & Williams suggest β = –0.25), as it might be for repeated clustering, then take, for example (Start beta: –0.25)&(Finish beta: –0.25)&(Num. increments: 1). You will also need to specify a factor for the SIMPROF groups, e.g. Add factor named: Flexbeta, which gives a Multi-plot (see next section) of the dendrogram and the line plot of ρ *vs.* β.

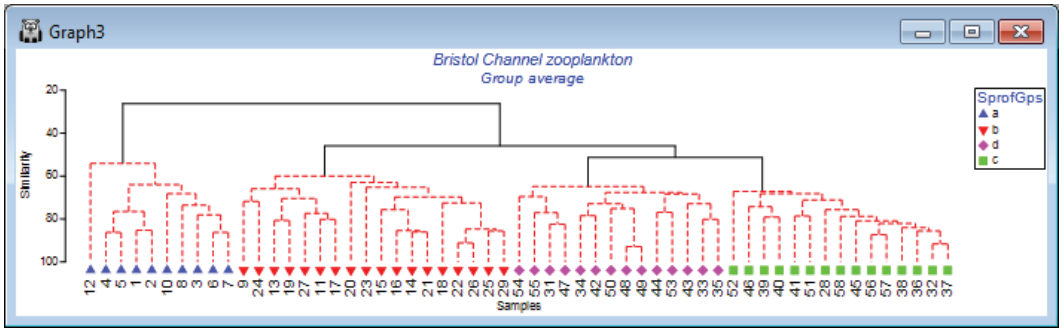
v7

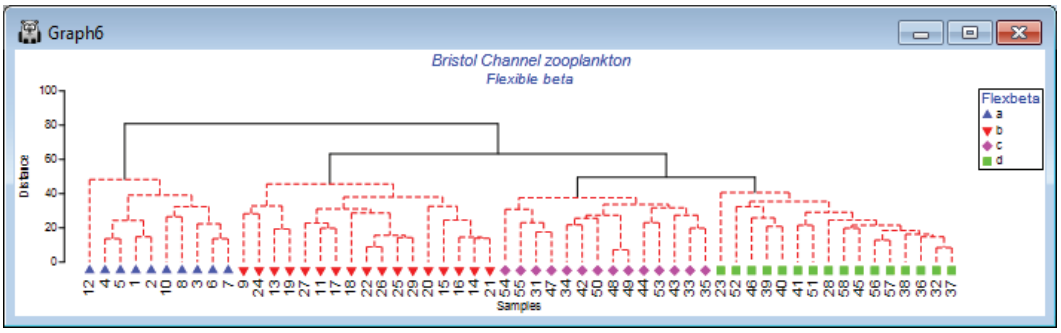


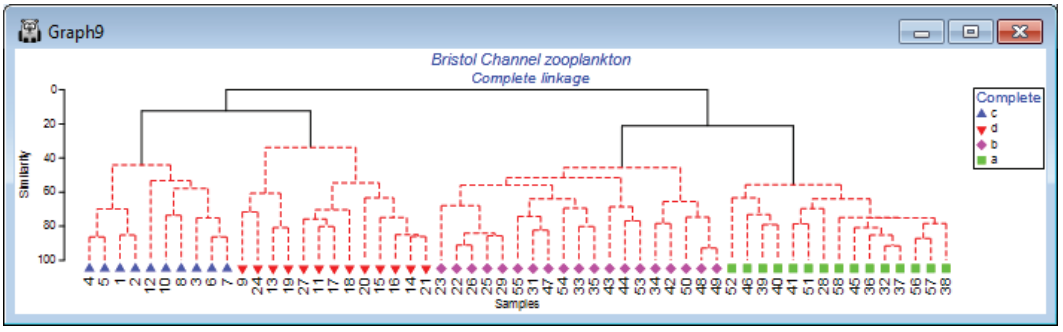
Single and complete linkage

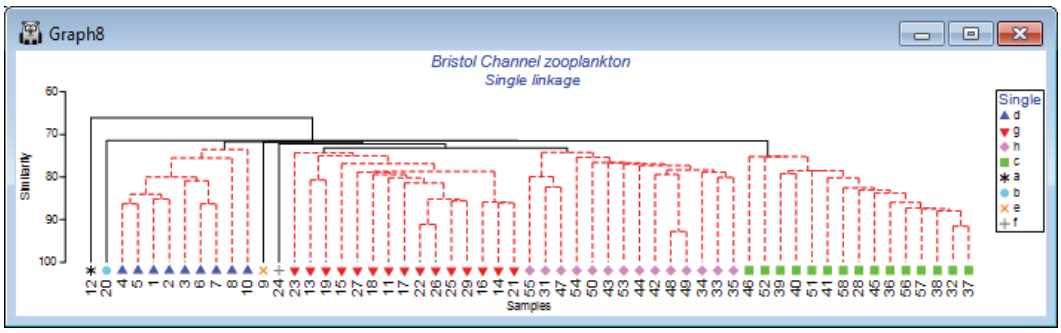
Now re-run **Analyse>Cluster>CLUSTER>**(Cluster mode•Single linkage), creating the SIMPROF factor *Single* and run again with (Cluster mode•Complete linkage), giving factor *Complete*. The respective cophenetic correlations ρ for the four linkage methods are: 0.797 (group average), 0.793 (flexible beta with β = –0.04), 0.722 (complete linkage) and 0.633 (single linkage). Though it can be difficult visually to compare dendrograms, careful rotation of plots and using **Graph>Sample Labels & Symbols** to put the respective SIMPROF group factors (*SprofGps*, *Flexbeta*, *Complete* and *Single*) onto the *x* axis as symbols shows that: Group average and Flexible beta differ only in the allocation of site 23 between the four SIMPROF groups (they are often similar but the flexible beta method is usually slightly inferior to group average); Complete linkage is similar in that four SIMPROF groups are also defined, though with a sub-cluster of 22, 23, 25, 26, 29 moving between two of these groups (to the detriment of the cophenetic correlation); and Single linkage is the only plot to look substantially different, with clear 'chaining' of samples and some singleton SIMPROF groups (sites 9, 12, 20, 24), with a clearly poorer fit (ρ = 0.63) to the similarity matrix. (It is often easier to visualise such changes in SIMPROF groups from differing cluster options by indicating those groups as symbols on an *MDS ordination*. For description of the latter see Section 8 – and for an example of the type of comparative MDS plots suggested see Fig. 3.10 of CiMC).

v7



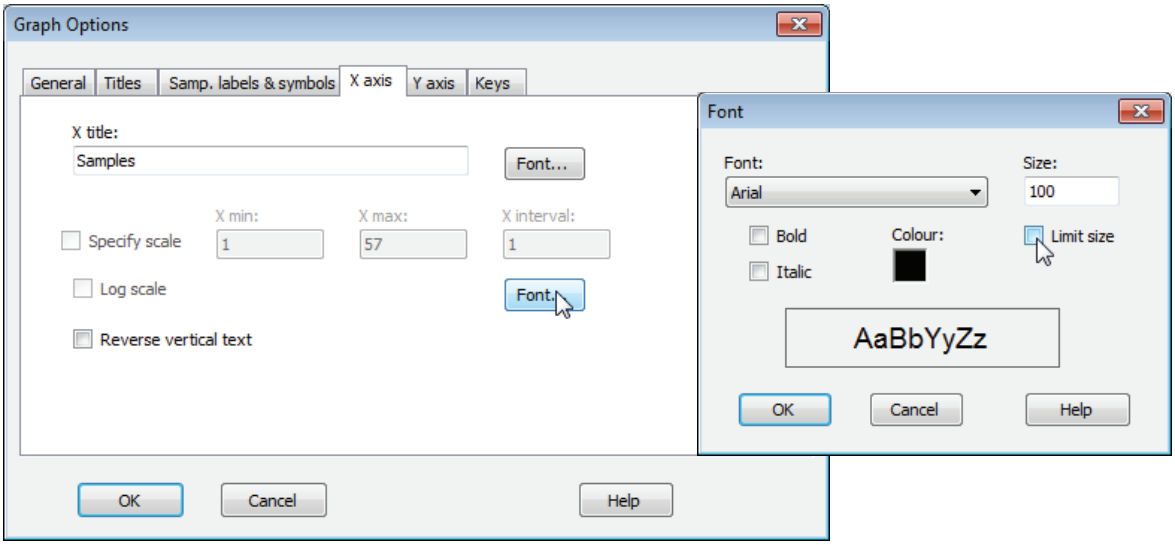






Limiting font size

Note that the plots above required a certain amount of juxtaposition of different font sizes for titles, axis titles, *x*-axis labels, keys etc, away from the default values (usually 100). Changing **Graph>General**>(Overall font scale: 100) is sometimes a good place to start, but you will see here that an increase (to 150 for example) does not always result in a font size change because, by default, there are upper size limits on much of the lettering, to avoid labels overwriting each other or parts of the plot. To override such a default, which has been carried out for the *x*-axis site labels (to increase their size almost to touching), click anywhere on the *x*-axis labels which throws you automatically into Graph Options on the X axis tab, select the relevant Font and untick the (✓Limit size) box. The other operations needed here were to re-order factor levels in keys and switch colours/symbols for some groups, by clicking on the key, using (Move>↓ or ↑) etc, as seen earlier in this section.



Binary divisive clustering

v7

Two new clustering methods are introduced towards the end of Chapter 3 in CiMC, the first still a hierarchical clustering method leading to a tree diagram, but a divisive rather than agglomerative algorithm in which all samples start off in a single group and are then split into two groups, each of those then further sub-divided into two, and so on until some stopping rule is activated. The sub-groups are not constrained to be of comparable sizes, in fact may sometimes be a split of *n* samples into a group of size *n*–1 and a singleton. In keeping with the principles embodied by the PRIMER package, the criterion which is maximised in making each split is the non-parametric ANOSIM *R* statistic of Section 9, used as a pure measure of group separation for a multivariate set of samples (and not in any way as a test statistic). *R* is essentially the difference between the averages of rank dissimilarities between two groups and averaged rank dissimilarity within those groups, suitably scaled so that it takes values up to +1 (*perfect* rank separation, in which all dissimilarities between the groups are larger than any dissimilarities within either group). After each binary division, the dissimilarities among samples within each new group are re-ranked, and used to maximise *R* in a further binary division. Even for quite modest sample sizes, evaluating *R* for all possible splits into two groups can be prohibitive, so a search algorithm is required and the number of random restarts of that process needs to be specified (default 10, but increase this if the routine runs quickly). A range of different stopping rules are allowed, which can be used in combination: a) a split which would produce a group of size *n* or less is never made (*n* specified); b) groups of size <*n* are never split (*n* specified); c) a split is not made if the largest *R* is less than a specified value; d) a group is never split if a SIMPROF test of its samples cannot reject the hypothesis of 'no structure' within that group – this is the least arbitrary and most natural of the stopping rules, a natural counterpart to the stopping rule for interpretation used for the agglomerative clustering described earlier.

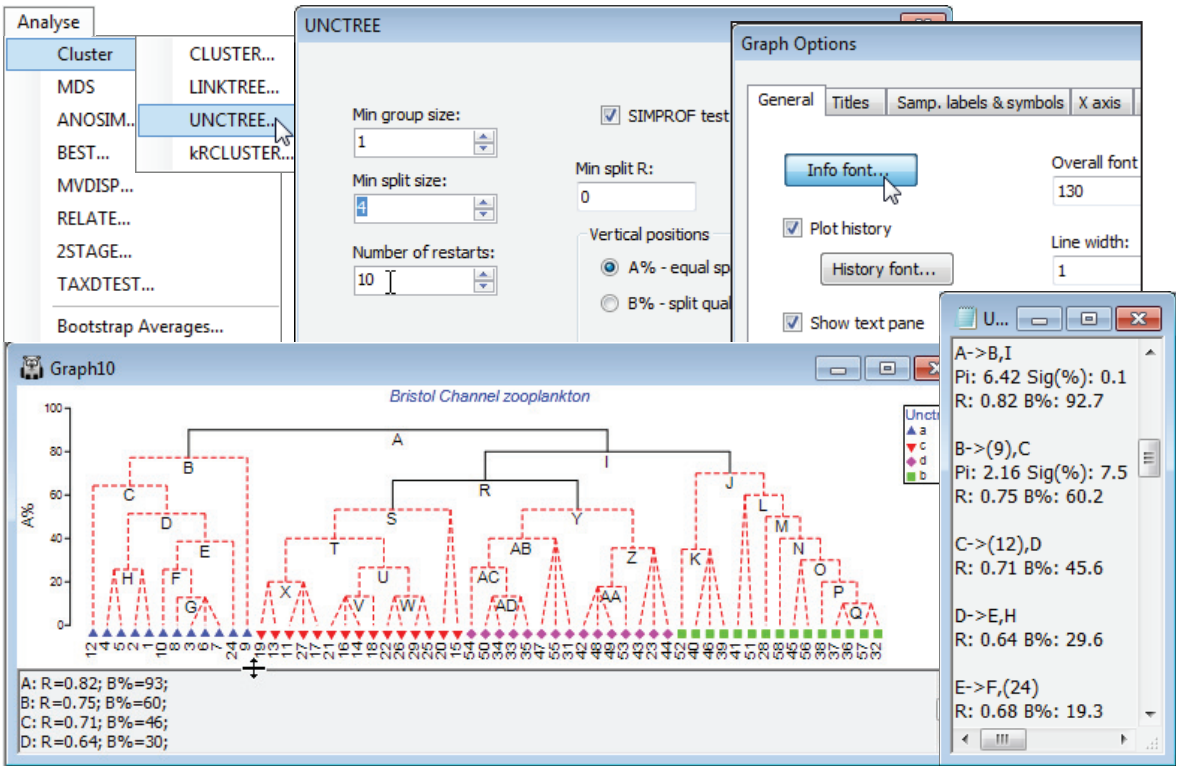
A parallel routine **Analyse>Cluster>LINKTREE** is described in Section 13 (called *linkage trees*), a constrained divisive clustering in which binary splits of, for example, biotic community samples are made in the same way (by maximising *R*), but only if an environmental variable can be found that takes a non-overlapping range of values in the two groups produced (a possible 'explanation' for that split therefore). In contrast, this new routine to PRIMER 7 is a completely *unconstrained tree*, accessed by **Analyse>Cluster>UNCTREE**: each sample is divided to maximise *R*, based only on the input resemblance matrix, e.g. the community similarities, without external constraints.

v7

UNCTREE options

v7

Using the Bristol Channel zooplankton workspace which should still be open, and with active sheet the Bray-Curtis similarities on fourth-root transformed data, run **Analyse>Cluster>UNCTREE>** (Min group size: 1)&(Min split size: 4)&(Number of restarts: 50)&(Min split R: 0)&(✓SIMPROF test)&(Vertical positions•A% – equal spaced) and take the defaults on the SIMPROF dialog, which is exactly that described for the CLUSTER routine above. Add factor name: Unctree, which holds the SIMPROF group labels which can then be compared with those for the previous dendrograms. Here, *Min group size: 1* imposes no constraint on how small a group can be, and *Min split R: 0* effectively takes out this stopping rule (max *R* will always be > 0), but *Min split size: 4* does come into play, so that once a group reduces to three samples it is not further divided. SIMPROF would have the ability to identify a group of three as heterogeneous – though the differences must be stark for it to do so – so there is no strong reason to impose this constraint. (However, SIMPROF cannot ever generate a significant result for a group of two; see CiMC, Chapter 3). The main guide here to interpretation will be the series of SIMPROF tests, with the parts of the tree drawn as red dashed lines again having no statistical support. In the absence of other strong information to the contrary, interpretation should thus be confined to the groups identified by the continuous black lines.



Text pane in tree plots

Note how each node is now lettered so that information about the *R* value for that split can be displayed in the *text pane* underneath the tree, and also in the accompanying results window (right). The lettering size in the tree can be increased (as it has been here) by **Graph>General>Info font**. The lettering order looks a little haphazard but this is only because the tree has been rotated exactly as seen for the earlier dendrograms – by clicking on the horizontal lines – in order to allow a better comparison with the previous agglomerative clustering. (In fact, the SIMPROF tests again result in only four groups, largely similar to those found before). The text pane can be scrolled and dragged down or up to smaller or larger heights but it but does not serve a particularly strong function here and could be turned off altogether by unchecking the (✓Show text pane) box on the **General** tab. It comes into its own for the companion routine of *constrained* divisive clustering (LINKTREE) seen in Section 13. There the text pane will list, for example, the inequalities on environmental variables which are capable of 'explaining' each lettered division of the biotic communities. Note also that, as for agglomerative clustering, details are given in the text-based results window (UNCTREE1) of the π statistic and its significance level from the SIMPROF test at each node for which the test is performed. For example, the initial split A into groups B and I is highly significant ($p<0.1\%$) but that at B, into the group C and the single sample 9, only achieves a non-significant level according to the criterion for continuation set in the SIMPROF dialog ($p<5\%$). Thus no more tests are carried out on the nodes further down this branch (C, D, E, …), as can be seen in the results window.

A% and *B%* *y*-axis scales

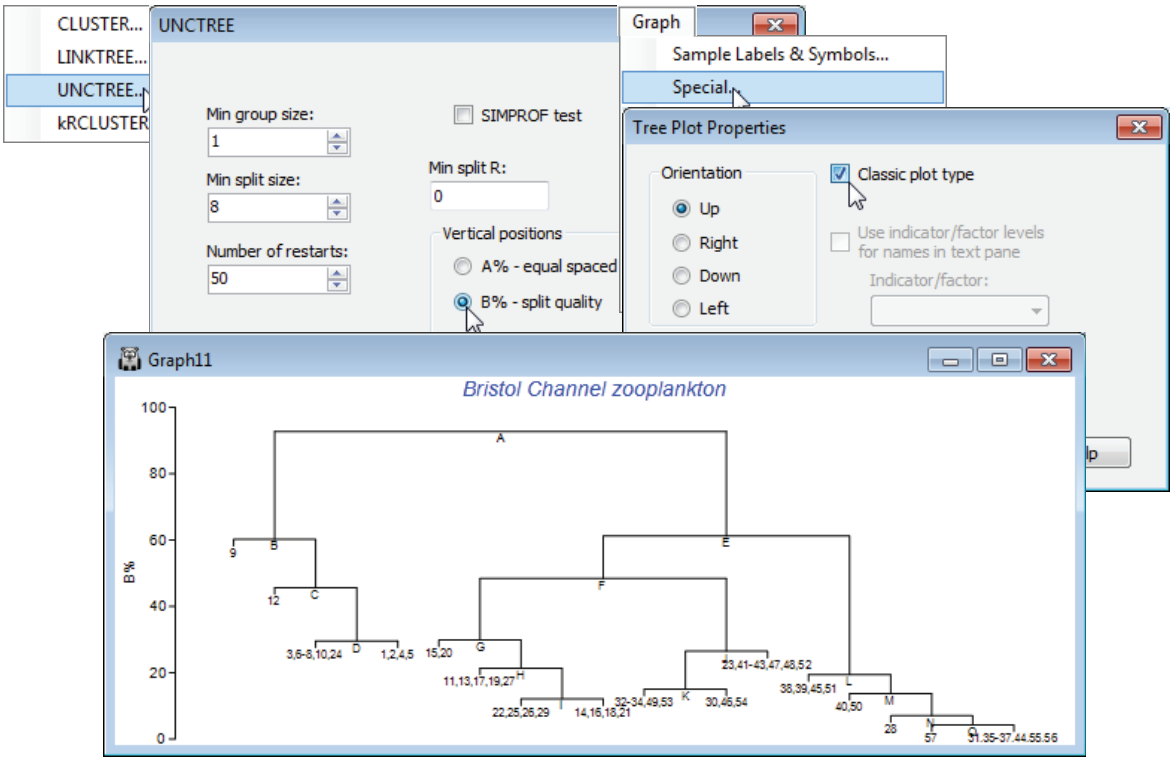
v7

The choice of *A%* as the *y*-axis scale above, evens out the spacing of steps down the binary tree in essentially arbitrary fashion, to give an uncluttered presentation – values of *A%* in different parts of the tree are not then quantitatively comparable (and a cophenetic correlation coefficient would make no sense). The alternative, originally described for the LINKTREE routine by Clarke KR, Somerfield PJ, Gorley RN 2008, *J Exp Mar Biol Ecol* 366: 56-69, is to take the (Vertical positions• B% – split quality) option in the UNCTREE dialog, in which average rank dissimilarities between groups on the original ranks (not re-ranked at each stage) are used to define a scale reflecting the magnitude of a division, in relation to the overall scale of variation (e.g. community change) across the full set of samples. The *B%* values are therefore comparable across different parts of the tree.

Special menu for divisive trees

v7

PRIMER 7 also provides a choice of representations of the tree structure, using either the *A%* or the *B%* axis scale. In general, the layout shown above is to be preferred, because the regular spacing of the sample axis allows non-numeric labels and/or symbols to be added, exactly as for a CLUSTER dendrogram. However, the Clarke *et al* 2008 paper used the tree layout shown below, referred to as *classic format*, an option from the **Special** menu, and this may still occasionally be found useful. Re-run the divisive clustering, this time with **Analyse>Cluster>UNCTREE**>(Min group size: 1) & (Min split size: 8) & (Number of restarts: 50) & (Min split R: 0) & (Vertical positions•B% – split quality), unchecking the (✓SIMPROF test) box this time. On the right-click menu, when over the plot, take **General** and uncheck (✓Show text pane), setting (Overall font scale: 140), then **Special**> (✓Classic plot type). The other options on this Tree Plot Properties dialog are not relevant here. In the standard plot mode they would allow the tree to be shown on its side or inverted (as seen earlier for the equivalent **Special** menu for a dendrogram from CLUSTER), and the greyed out option is only applicable to the constrained form of this divisive clustering (LINKTREE), where explanatory variable names (e.g. environmental) are given as inequalities in the text pane, and it is convenient to expand or abbreviate the names using an indicator defined on those variables, see Section 13.



Two features are apparent from this plot. Firstly, the use of *B%* scaling on either type of plot does show (as a dendrogram would) that the divisions lettered A , E, F are major divisions between the clusters, in relation to the subdivisions of those groups, e.g. at H, J, L etc at much lower levels on the *y*-axis scale; this fact is missing with the equi-stepped *A%* scale. However, there is the potential for *reversals* when using *B%* scaling (sub-cluster divisions returning higher values of *B* than their parent split) especially in the constrained form of this clustering (Section 13). Secondly, note that labels on the 'classic' plot have to be sample numbers, exploiting number ranges to keep the plot tolerably neat (text labels would be impossible), which is highly confusing here when the sample sites are actually labelled 1-29, 31-58 (site 30 not sampled), but the sample numbers will be 1-57!

Flat-form clustering

v7

Another new introduction in PRIMER 7 is a form of non-hierarchical (*flat*) clustering, the analogue of the *k*-means method in traditional cluster analysis. The latter is a widely-used technique based on Euclidean distances in the *variable space* of the original data matrix, seeking to form an optimal division of samples into a specified number of groups (*k*), minimising the within-group sums of squares about the *k* group 'centres' (termed *centroids*) in that high-dimensional variable space. However, in that form, it is quite inappropriate for typical species matrices, for which Euclidean distances or their squares (whether on normalised variables or not) are a poor measure of dissimilarity among samples, as discussed in Section 5 and in more detail in Chapters 2 and 16 of CiMC. What is required here, to be consistent with the rest of the PRIMER package (and the hierarchical methods previously described) is a technique which applies to any dissimilarity coefficient, and in particular, those suitable for species data (e.g. Bray-Curtis). By analogy with *k*-means, the concept of *k*-*R* clustering is introduced towards the end of Chapter 3 of CiMC, in which the k groups are chosen to maximise the global ANOSIM *R* statistic (as it would be calculated for an ANOSIM test of the *k* groups involved). Again, the use of *R* here has nothing to do with hypothesis testing; it is its usefulness as a completely general measure of separation of defined groups of samples, based only on the ranks of the dissimilarity matrix – the same numbers, however that dissimilarity is defined – which is being exploited. Above, we used the idea of maximising *R* for a division of the samples into two groups; here the **kRCLUSTER** routine simply generalises that to maximising *R* calculated over *k* groups. It again involves a demanding iterative search, with user choice of the number of random restarts (again the current default is 10 but try more if the process runs quickly).

v7

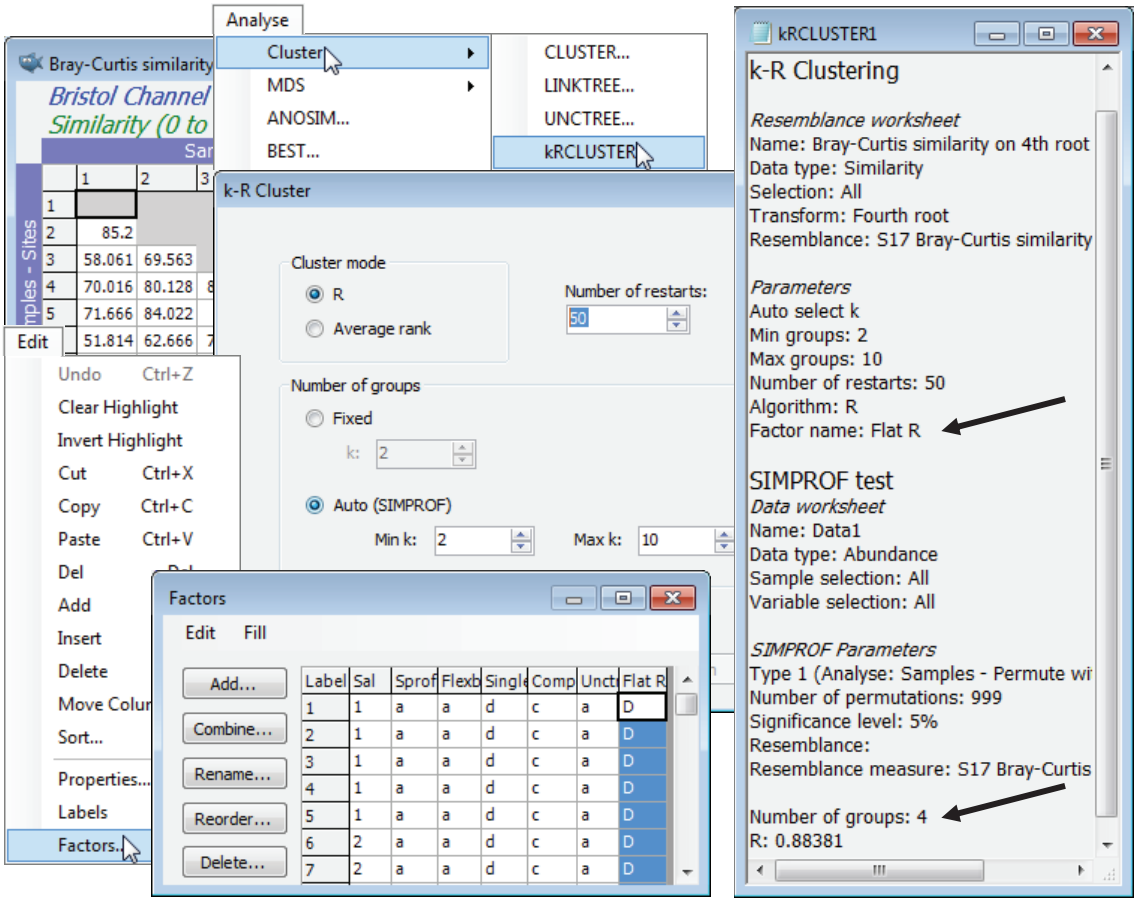
A perceived drawback of the *k*-means approach is that *k* must be specified in advance. There may be situations in which a pre-fixed number of groups is required but, more likely, it would be useful to determine the 'best *k*' from a range of values, in some well-defined sense. SIMPROF tests can be exploited here also, to provide a possible stopping rule. The k-R Cluster dialog asks for min and max *k* values to try, and starting with (say) the default min *k* of 2, finds the optimal division into two groups and tests those groups, with SIMPROF, for evidence of within-group structure. So far, these groups and the tests will be exactly those of the unconstrained binary divisive (UNCTREE) routine, above. But these groups are not then further subdivided – this is not a hierarchical process. If at least one SIMPROF test is significant then these groups are thrown away, and the procedure starts again with the full set of samples and attempts to find an optimal *k*=3 group solution. These groups are again tested with SIMPROF, and if any of the three tests is significant, a *k*=4 solution is sought on the full set of samples, etc. The procedure stops either when the specified max *k* (default 10) has been explored or when all SIMPROF tests for the current *k* are not significant (i.e. there is no statistical evidence of structure at a finer-scale than this *k*-group partition). **kRCLUSTER** will request a factor name to define that grouping; note that it is a single factor holding only the solution for the (optimum or maximum) *k*-value at which the procedure terminates. A tree diagram cannot, of course, be plotted, since there is no hierarchy. In fact, the reason for exploring *flat* clustering of this type is to avoid the inflexibility, in hierarchical methods, of samples being unable to 'change their allegiance' – once in a specific group, a sample remains in a subset or superset of that group.

v7

A final choice on the k-R Cluster dialog is between (Cluster mode•R), which is precisely the rank-based algorithm described above, and (Cluster mode•Average rank), which is a subtle variation bearing some analogy with group average linkage (an idea met in agglomerative clustering but here still used to produce a *flat* clustering). The last page of Chapter 3 of CiMC explains this variation, which (though not using *R* as such) is still a function only of the ranks of the original resemblance matrix. In practice, the two flat-clustering modes should produce rather similar solutions.

v7

Again on the Bristol Channel zooplankton data, whose workspace should still be open, with the active sheet as the Bray-Curtis similarity matrix based on fourth-root transformed densities, take **Analyse>Cluster>kRCLUSTER>**(Cluster mode•R) & ((Number of groups•Auto (SIMPROF))> (Min k: 2) & (Max k: 10)) & (Number of restarts: 50), and with defaults taken on the SIMPROF options dialog, and specifying factor for the optimal grouping of Flat R**.** The results are inevitably rather minimal in this case: the results window gives the optimal number of groups again as *k*=4 (with *R*=0.884), and **Edit>Factors** will show the Flat R grouping. You may like to run the routine again with (Cluster mode•Average rank), which results in the same *k*=4 groups here, though the factor sheet shows that the order of assignment of letters A, B, C, D to the 4 groups may differ. This is an inevitable result of the random search procedure, even when the same options are taken.



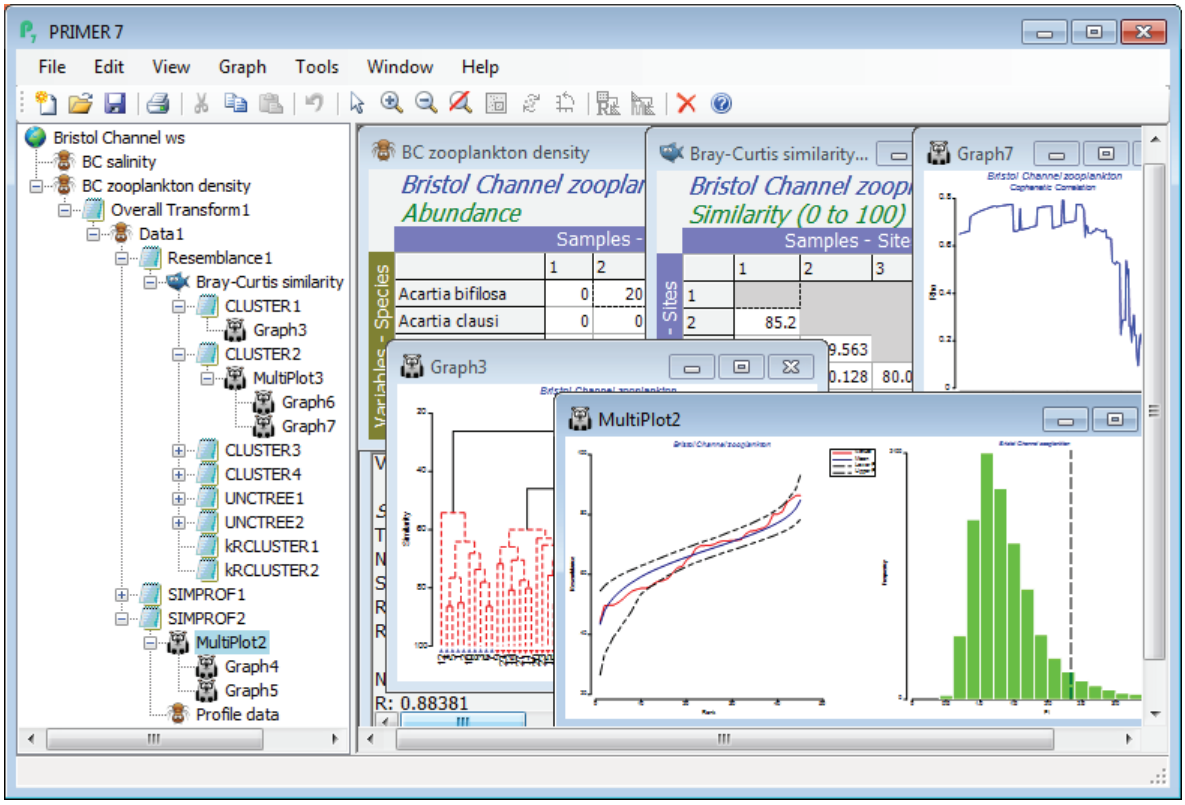
In fact, much the best way of comparing the results of the differing clustering methods of this section is seen for these data in Fig. 3.10 of CiMC, namely on three copies of the same non-metric MDS ordination of the 57 samples. See Section 8 for running MDS ordinations, so this example will not be pursued further here (but you might like to return to these data after tackling Section 8 and reproduce a larger version of Fig. 3.10, covering all the variations of clustering methods you have generated in this section, so **Save Workspace As**>File name: Bristol Channel ws). In Fig. 3.10, the differing SIMPROF group factors *SprofGps*, *Unctree* and *Flat R* – for the hierarchical agglomerative (group average), divisive and flat clusters – are plotted as symbols, and relettered consistently, since essentially the same four main groups result from these very different clustering techniques. The minor differences between methods are clear: they just concern allocation of a few sites, which tend to be intermediate between the main groups – the treatment of sites 9, 23 and 24 is all that distinguishes them.

This is exactly what one might wish for in drawing solidly-based inference of clustering structure – a stability to the choice of method. It is relevant here that the same transformation and (especially) similarity matrix was used for all methods. Major differences in groupings would be expected to arise from using different pretreatments or dissimilarity definitions, e.g. comparing SIMPROF groups from agglomerative clusters, using Bray-Curtis on fourth-root transformed densities, with SIMPROF groups from a method closer to traditional *k*-means clustering (normalised species data, with resemblances calculated using Euclidean distance, and analysed by the Average rank cluster mode of the above *k-R* clustering). This has rather little to do with choice of clustering method but everything to do with what is understood by similarity of samples in the high-dimensional species space. This is a recurring theme in CiMC: the major differences between ordination methods such as PCA (Section 12) and nMDS (Section 8) usually has much less to do with the different way the methods try to view high-d data in low-d space, but much more to do with how those methods choose to define ‘distances’ in that high-d space at the outset (PCA by Euclidean distance, nMDS often by a species-based community measure from the Bray-Curtis family).

7. Managing the workspace and plotting (*Window*, *File*, *View*, *Multi Plot*, *Plots*)

Explorer tree

It will have been obvious from performing the worked examples of the previous sections that the structure of all calculations in the current workspace is stored and displayed in logical fashion, on the left of the PRIMER desktop, in the *Explorer tree*. This lists data sheet names, and any derived constructions (such as further data sheets, resemblance matrices or plot windows), via the intermediate Results windows, which are simple text windows giving details of the options taken in moving down a branch from one sheet to an ensuing window, in addition to listing textual results of the operation. This allows the user to manage the workspace, finding and activating only windows which are needed, and saving the entire workspace structure in a single operation, and in a single file, for later retrieval. For example, if you have been analysing the data of the previous section, the PRIMER desktop should now look something like:



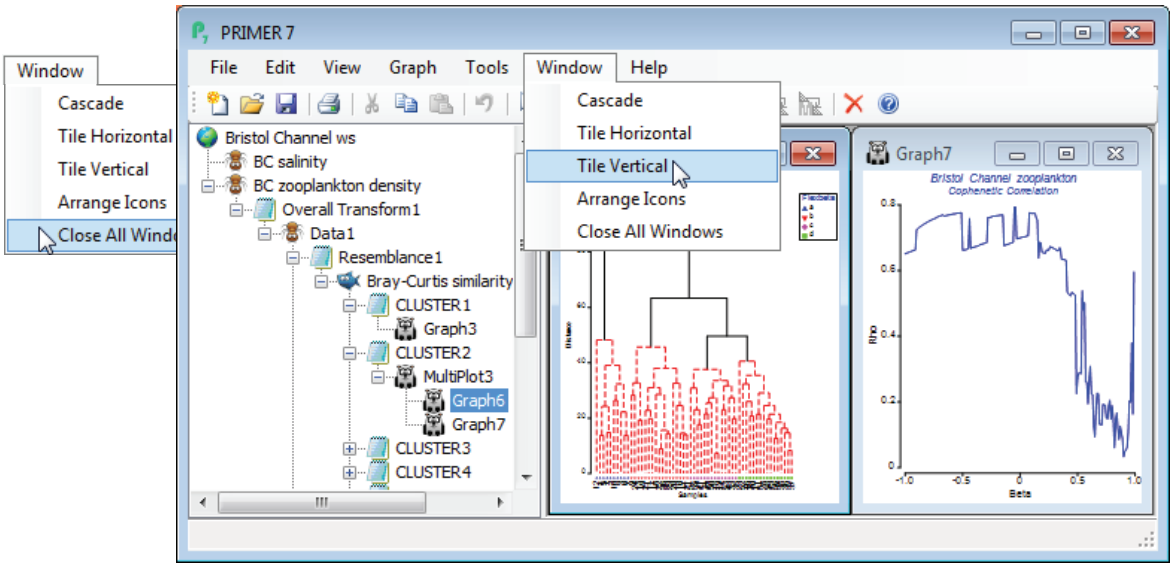
Forward and backward propagation

The Explorer tree not only makes it easy to navigate the sequence of steps taken in an analysis but also, importantly, reflects the program's internal knowledge of the inter-relationships among data, resemblance sheets and plot windows. It uses this structure to select sensible defaults, and even occasionally to reach *back* and find a data matrix higher up the same branch, needed for a specific operation. (An example of this has been seen in running SIMPROF from the various clustering methods – these routines are all launched with a resemblance matrix as the active sheet, but the tests require permutation of rows or columns of the preceding data matrix).The Explorer tree is also used for forward and backward propagation of factor/indicator information, along a single branch of the tree. A factor's properties, such as symbol types for each level, are naturally passed down a branch as new data sheets or plots are formed from the data sheet for which that factor was created. The reverse is also true: definition of new factors, produced from an editing step on a plot, will typically be passed back to the data matrix at the head of that branch, and then down other branches leading from the original sheet. But such information is not passed from one branch to a distinct branch (factors can be retrieved from distinct branches by **Edit>Factors>Import**), and there are also natural blocks to propagation. For example, if a **Tools>Average** (or **Sum** or **Merge)** operation – see Section 11 – is carried out part way down a branch, PRIMER 7 (unlike PRIMER 6) will now place the resulting averaged sheet on the same branch as its parent matrix, and all factors will be passed forward to the condensed sheet (though some factors may have undefined entries if the averaging has been over levels which differ for those factors). However, changes now to factors of the condensed matrix clearly cannot sensibly be backward propagated to the original larger sheet.

v7

Closing, redisplaying & tiling windows

From the Explorer tree for the above illustration, you can see that workspace has been saved as Bristol Channel ws, with every row below this representing a window in the workspace, whether open or not. It is important to appreciate that whilst you can close down windows – individually by the usual icon (top right), or all of them at once with **Window>Close All Windows** – they will remain in the workspace, and can simply be re-displayed by clicking on their name (or icon) in the Explorer tree. The option to tile or cascade windows may also be useful here, so that a common way of tidying up the display area (right panel of the PRIMER desktop) is to close all windows, click on the Explorer tree names for the ones you want to re-display and then take **Window>Tile Horizontal** or **Tile Vertical** or **Cascade**.

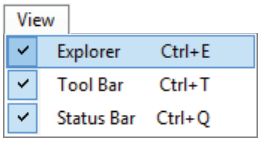


Minimising windows

For consistency with much earlier PRIMER versions and other Windows software, there is also an option to minimise a window to the bottom of the PRIMER desktop, by clicking on the minimising icon , but this is no longer necessary or useful with more recent versions of PRIMER, it being easier simply to close the window and retrieve it from the Explorer tree when required.

View menu

There will be situations in which, in order to display as much of a plot or datasheet on screen as possible, it is desirable to temporarily hide some of the other features, to widen and heighten the display area. In standard Windows fashion, the **View** menu allows three features to be toggled off and on again: ✓**Explorer** (the left panel), ✓**Tool Bar** (icons in the top row) and ✓**Status Bar** (bottom row, used to display position of the cursor in a data matrix, progress of a calculation etc).



Understanding the Explorer tree

Understanding the Explorer tree, however, is the key to managing your work pattern in PRIMER. In the above workspace, click on the successive entries. After the workspace name (Bristol Channel ws) two data matrices have clearly been opened into the workspace, BC salinity (a repeat of the salinity information for each site but in the form of a separate environmental matrix, rather than a factor of the zooplankton array – this will be useful later when displaying bubble plots of salinity on an MDS plot), and the community data BC zooplankton density. A fourth-root transform has been taken, indicated in the Overall Transform1 results window, giving Data1. The similarity calculation follows: Resemblance1 shows that this was Bray-Curtis, calculated for all samples using all species, to give the triangular resemblance sheet – by default this would be *Resem1* but has been renamed (e.g. by clicking twice on the name) to Bray-Curtis similarity on 4th root. Renaming important windows in this way is always a useful aid to navigating around a large workspace (and more of this should probably have been carried out in the current example!). The agglomerative (Group Average) cluster analysis generates the CLUSTER1 results window and the dendrogram plot, Graph3, the number indicating that two other graphs were produced before this (in an analysis from Section 6, under SIMPROF1 lower down the Explorer tree, on a

branch stemming back to the transformed *Data1* sheet rather than from the Bray-Curtis similarity matrix). Note that all these steps, which create a derived sheet or plot file, possess an intervening results window. This is what distinguishes operations on the **Pre-treatment**, **Analyse**, **Plots** and **Tools** menus (and **PERMANOVA+**, if installed) from those under **File**, **Edit** or **Select**. The latter do not produce new, derived sheets or intervening results windows, but make changes (temporary or permanent) to the currently active sheet.

Rolling up branches of the tree

v7

In the display of the Explorer tree at the beginning of this section, the SIMPROF1 window was prefaced by the *rolled-up* icon, and it is necessary to click on this to expand the branch below this point, replacing the rolled-up icon with one indicating that the branch is now *rolled-out* (in this case to reveal a *multi-plot* and the two plots, *Graph1* and *Graph2*, which have been collected together under this multiple plot construction, new to PRIMER 7, see below). A second click will reverse this operation, rolling up the branch below that point. This may be a useful way of keeping secondary analysis strands available in the workspace, but without allowing their detailed steps to clutter up the main sequence of analyses displayed in the Explorer tree. Any windows which have been closed in the PRIMER desktop (not deleted from the workspace), e.g. by an individual or a full **Window>Close All Windows** operation, at the time the workspace is saved, will appear only under a rolled-up entry when the workspace is re-opened. As we saw earlier (in Section 1, under the Saving, closing & opening a workspace heading) there is a new option to open a workspace with all its branches in rolled-up form, irrespective of how the workspace was saved – thus opening the workspace more quickly – requiring a succession of roll-outs to display individual sheets.

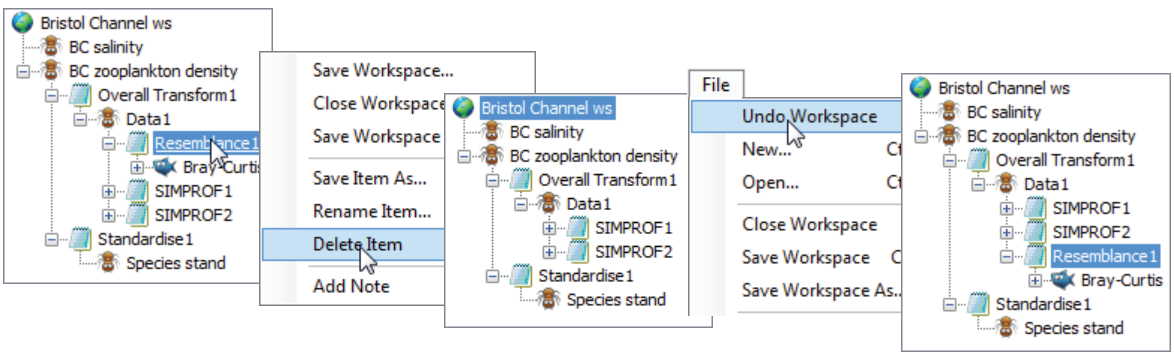
Renaming or deleting items in a workspace

Met briefly in Section 1 and again above, renaming or deleting windows in the Explorer tree is an essential part of keeping a workspace navigable and understandable, especially since workspaces for many real analyses can become voluminous! Renaming is accomplished in one of three ways: as mentioned above, by clicking twice (slowly) on the entry name and typing directly into its box; by taking **File>Rename Data** (this changes to **Rename Results**, **Rename Resem** etc, depending on the entry type); or by right-clicking when over the Explorer tree to obtain a 'floating' menu which includes a general **Rename Item** operation. If part of an analysis is wrong or unhelpful, and you wish to delete it altogether from the workspace, a similar pair of options exists: click on the results name or icon and take **File>Delete Results**, or from the floating (right-click) menu take **Delete Item**. This results window, and all items below it on the same branch (its *derived* windows) will be erased from the workspace. You are prompted with the entry name to make sure that this really is the window (and derived windows) that you want to delete.

Undo in the Explorer tree to reinstate or re-order

v7

Note that in PRIMER 7 such a deletion is now a reversible operation, with **File>Undo Workspace**, which can be operated repeatedly to back-track through many successive **Delete** and/or **Rename** steps on the **File>** menu (though not, of course, **Save**, **Close**, **New** or **Open** operations since they are either easily back-tracked in other ways or are patently irreversible – such as saving to external files, or closing a workspace and ignoring the warning to save it first). You might like to try this out on the current Bristol Channel zooplankton workspace (save it first before you experiment!). Note what happens on reinstatement of branches or terminal windows (often plots) after deletion: they are added back, as might be expected, to the end of the stack of items at the same branch level, rather than the precise position from which they were removed (of course they retain exactly the same hierarchical position in the Explorer tree structure). One by-product of repeated deletion and reinstatement could thus be a limited ability to re-arrange the main strands of an analysis or the order of duplicated plots (e.g. a range of bubble plots, see next section) within the Explorer tree.



Saving plots

The name of a window can also be changed as part of the process of saving it, as a file external to the workspace: if the file is given a different name in the act of saving it, then that new name will also replace the existing entry in the Explorer tree. However, there is not the same need to save an individual data sheet, resemblance matrix, plot etc, that there was in the early versions of PRIMER (e.g. v5) because, as we have seen many times now, the full workspace can be saved in a single **Save Workspace As** operation, and this is a convenient way to pass data analyses to other users of PRIMER, for example. Nonetheless, there may be occasions when data (or resemblance) matrices need to be saved in an export format, as Excel *.xls(x) or text format *.txt files, or in internal v7 or v6 binary format – note that the v6 and v7 binary data formats differ – with extensions *.pri (or *.sid), perhaps so that they can be opened in a different workspace. More commonly, individual windows which are saved externally to the software are likely to be plot files.

From the current Bristol Channel zooplankton workspace, try saving the dendrogram Graph3, for group average agglomerative clustering, by highlighting it in the Explorer tree and taking **File> Save Graph As>**(Filename: BC zoo UPGMA) & (Save as type: PRIMER Plot Files (.ppl)). This *.ppl extension denotes PRIMER 7's internal binary format for graphics, and it too differs in v7 and v6 – if a v6 format plot file is required (for those plot types existing in the earlier version) then it should be explicitly chosen by (Save as type: PRIMER 6 Plot Files (.ppl)). The main use of *.ppl format files is likely to be in passing plots to other PRIMER users, so note that, whilst PRIMER 7 will read the v6 *.ppl format (*forward compatibility*), PRIMER 6 will not read v7 *.ppl files (no *backward compatibility*), hence the need to save explicitly in v6 format, on the rare occasion where this might be required. (Neither format can be accessed by any other software – even PRIMER v5). Demonstrate such transfer by launching a parallel run of PRIMER 7 (e.g. on the Windows 7 or 8 desktop, double-click on the desktop icon or right-click on the task bar icon and select PRIMER 7 again), to generate a second PRIMER desktop with an empty workspace. Open this newly created plot file into it, with **File>Open>**(File name: BC zoo UPGMA). You will see that, in spite of the link to its original data and resemblance matrix being cut, the plot is complete in itself, and still capable of being modified. It even holds with it the background information on factors, inherited from the data file, so all the changes seen in Section 6 can be implemented: not just resizing, titling, suppressing the key or history box, zooming, condensing, rotating, etc, but also changing displayed symbols to a different factor. It is important to realise that any changes made here will in no way affect the same plot held in the first PRIMER workspace. There no dynamic linking of any sort between different workspaces or between workspaces and files external to them; only an unlinked copy of any file is ever opened in, or saved from, PRIMER. It should also be emphasised that PRIMER 6 and PRIMER 7 can both be running at the same time and will not interfere with each other, so if both are installed you may wish to try saving the BC zoo UPGMA dendrogram again from v7, this time with (Save as type: PRIMER 6 Plot Files (.ppl)), and re-opening it in v6.

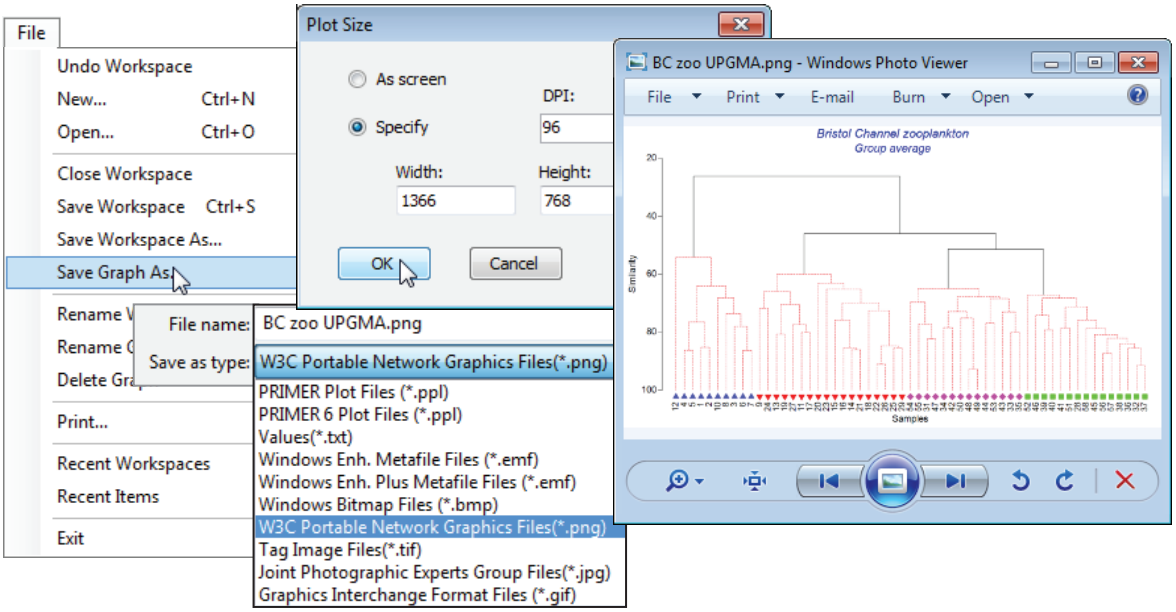
Vector *vs.* pixel plots

Closing the second PRIMER desktop and returning to the original Bristol Channel ws workspace, note the other options for saving a dendrogram or any plot file with **File>Save Graph As>**(Save as type:). The vector format Windows Enhanced Metafile (*.emf) will usually be the best option for exporting graphics from PRIMER into other applications, for fine tuning of title or key placement etc, in graphics presentation software. We saw earlier that **Edit>Copy**, when the active window is a plot, takes this vector format to the Windows clipboard, from where it can be pasted, for example, into Powerpoint. When Ungrouped it will be converted to a Microsoft drawing object and the lines, symbols, text boxes etc which make up all plots can be subsequently modified, as appropriate.

In contrast, the other static plot output options from PRIMER all produce bitmap (i.e. pixel-based) files: *.bmp, *.png, *.tif , *.jpg and *.gif formats. Subsequent modification options are then rather limited. However, if the plot can be put into a satisfactory finalised form using the manipulations available within PRIMER, then high-quality output is certainly possible through the bitmap route. Saving the plot in one of these formats allows specification of the resolution, e.g. (Plot Size•As screen) or (Plot Size•Specify), with specifications being, for example, (Width: 1024) & (Height: 768), which give width and height of the image in pixels, and also specification of dots per inch, e.g. (DPI: 96). These files will generally be much larger than for vector plots.

v7

v7 A new feature in PRIMER 7 is the ability to output some plots in dynamic form, in cases where this is appropriate, via video format *.mp4 or animated *.gif files. We shall see such graphs in the next section, e.g. in 2- or 3-d animations of MDS iterations, temporal patterns and rotated 3-d plots.



Saving graph values

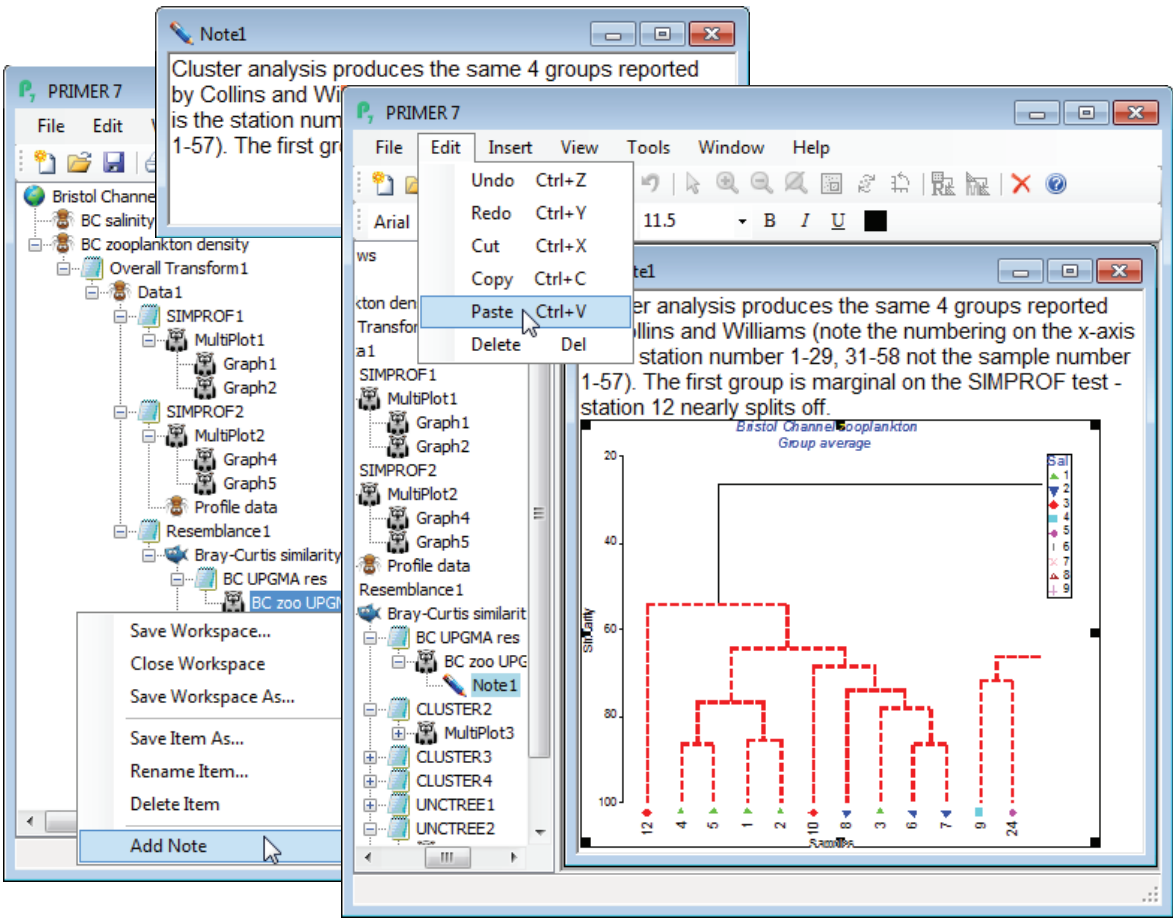
Certain graphs, such as MDS ordinations (Section 8) or Cluster dendrograms, can be validly rotated in an infinity of ways (effectively), after the results window is generated, perhaps to align them better with a previous run under different transformation or coefficient choice. The plot is always saved in its currently rotated state, naturally, but these will not then correspond to the co-ordinate positions of ordination points, for example, which are listed in the results window. In order to make available the current ordination co-ordinates, or in the case of a dendrogram the ordering of the samples on the *x*-axis under the current rotation, an option to Save Graph Values is provided. This can be run in two ways, by **File>Save Graph As**>(Save as type: Values (*.txt)) or more directly by **File>Save Graph Values As**. The end result in both cases is a text file containing either *x,y* or *x,y,z* co-ordinate points for an ordination (each point to a line and tab separated within a line), or a list of the current order of samples in the dendrogram (each sample label to a line).

Saving results

When they are active, results windows can be saved in just the same way, e.g. on CLUSTER1, **File>Save Results As>**(File name: BC UPGMA res) & (Save as type: Rich Text Files (*.rtf)) will save the individual clustering steps and associated SIMPROF tests to a file in rich-text format. The latter preserves variations in font size and use of italics when viewing the results window in Word, for example. The alternative is to Save as type: Text Files (*.txt) which outputs all text in a fixed size Courier font. The decision as to which option to select depends mainly on whether the text file will then be viewed in Word (use *.rtf) or in a simple text editor such as Notepad (use *.txt).

Adding notes

It is not permitted to edit directly the information in a results window. This tells you what operation or analysis was actually carried out, and what the outcome was, and should remain immutable, to avoid confusion if the workspace is revisited later. Naturally, you can highlight, then **Edit>Copy** (or Ctrl-C) results content to the Windows clipboard, and paste the information to an external text file, or Word or Excel file (tabular results from MDS, ANOSIM etc, see Sections 8 & 9, will map into appropriate Excel columns to allow simple editing and entry to other software – or indeed back into PRIMER). However, if you need to annotate the PRIMER session within the workspace, e.g. commenting on analysis steps or results, this can be achieved by **Add Note**, selected from the menu which appears when you right click on any item in the Explorer tree. A blank Note window is opened for typing, and is displayed in the Explorer tree on a branch leading from the originally clicked item, which could just be the workspace name, in which case the Note will appear at the bottom of the tree – a convenient place to put 'read-me' information. Text can be pasted into the note via the clipboard (**Edit>Paste** or Ctrl-V), from outside or from elsewhere in the PRIMER session (e.g. from a results window or information copied from the **Edit>Properties**>Description box, etc). You can even copy and paste whole graphs or highlighted portions of data sheets into a note window, so a note-form summary of the main features of the analysis can be held within the workspace (though lack of formatting usually makes this an intermediate step). Note windows can be renamed, deleted and saved, as with any other Explorer tree item, the save operation again involving a choice of *.txt or *.rtf formats (*.rtf is needed to preserve any plots in the output file).



Printing results and graphs

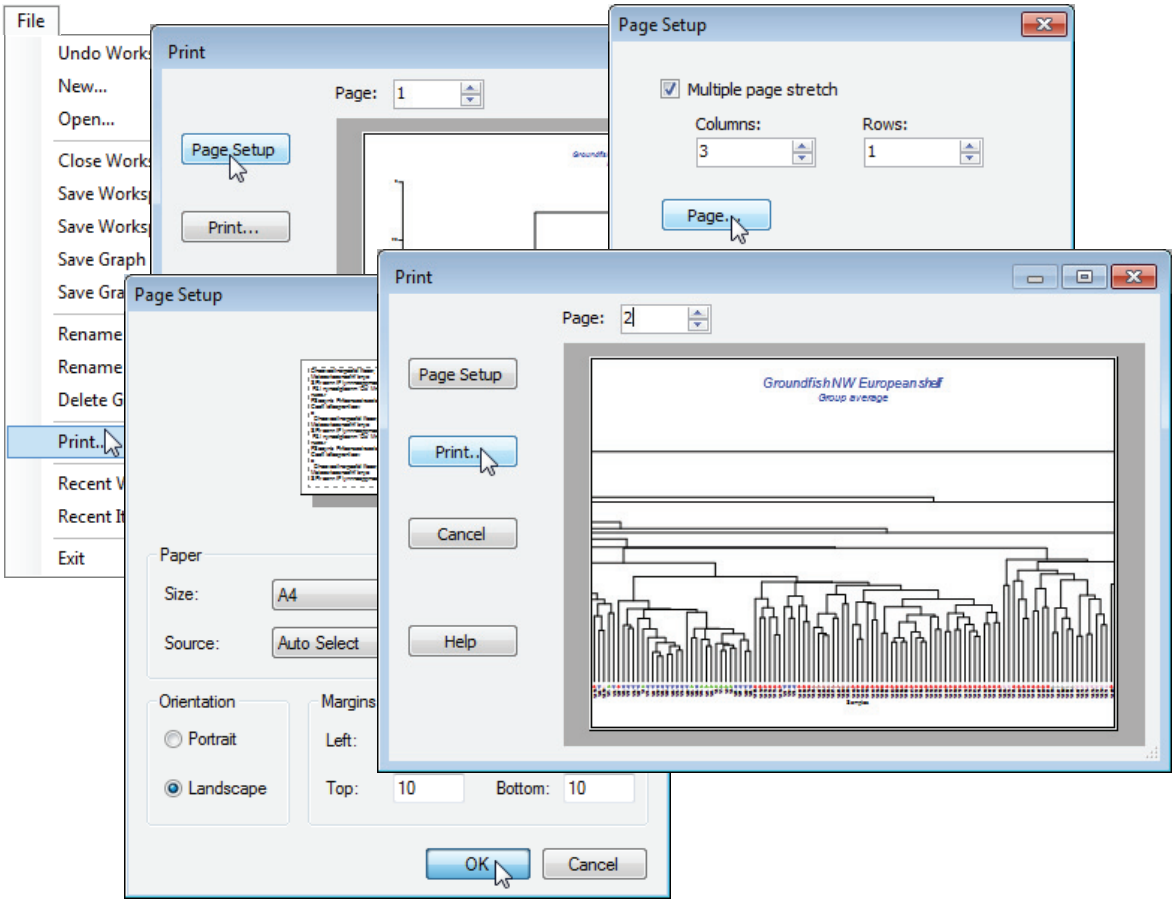
Direct printing from PRIMER is also possible for analysis endpoints such as results windows, all graphics windows, and notes (but not data sheets or resemblance matrices, which are generally too large and unwieldy for easy printing – selections of them are best saved to Excel or other software with capacity to size rows and columns into printable form). On plots, results or notes, **File>Print>Print** will take the default options and send you to the standard Windows dialog for selection of printer, any printer preferences, etc. However, for plots only, there is a PRIMER-specific option which can be taken prior to the final **Print** instruction, accessed by **Page Setup** after selection of **File>Print**. This allows a plot to be spread, horizontally and/or vertically, over multiple pages. It can sometimes be very useful in reading a cluster dendrogram based on many samples (thus printed over multiple horizontal pages) or in viewing a Shade Plot, perhaps from **Wizards>Matrix display** (Section 10), which can use a (✓Multiple page stretch) of the plot both horizontally and vertically. Also on the Page Setup dialog box, the **Page** button gives an alternative means of implementing simple printing choices of •Portrait or •Landscape, paper size and source, and margin sizes.

v7

The facility to stretch a plot over multiple printed pages is best illustrated for one of the previous data sets so, leaving the Bristol Channel ws workspace open, launch another run of PRIMER and, if the workspace from the C:\Examples v7\Europe groundfish directory is available (last saved in Section 6), take **File>Open**>(File name: Groundfish ws) and click on the dendrogram, Graph2, to make that the active window – or re-run that cluster analysis of 277 samples. By **File>Print>Page Setup>**(✓Multiple page stretch)>(Columns: 3) & (Rows: 1)>**Page>**(Orientation: Landscape)>**OK**, the viewing window in the Print dialog box now shows the left side of the dendrogram as (Page: 1), and changing that to (Page: 2) displays the centre and (Page: 3) the right side, with some overlap to aid the physical pasting of the three printed pages which now result from **Print**. You will also find that an optimal printing will need to greatly reduce font and symbol sizes for all elements of the plot – in fact for most, if not all, of the font options (including Overall font scale) on the **General**, **Titles**, **X axis**, **Y axis** and **Keys** tabs of the Graph Options dialog and also the symbols plotted in the key, whose sizes are controlled from the (Size:) option under **Samp. labels & symbols**.

v7

File>Save Workspace the Groundfish ws and close down this second PRIMER desktop by the **File>Exit** menu item, leaving open the Bristol Channel ws workspace.



Automatic creation of multi-plots

v7

The concept of a *Multi-plot* is a new feature in PRIMER 7. This construction has already been met in Section 4 where histograms of all selected environmental variables, using frequencies calculated over the samples, were presented in a single multi-plot window with component graphs consisting of the individual histograms for each variable (using **Plots>Histogram Plot**). You may have noted it again in the Explorer tree for the above Bristol Channel zooplankton analyses – see the PRIMER desktop in the first screen shot of this section – where a direct SIMPROF run (on one of the groups of samples identified by the cluster analysis) generated two plots: the similarity profiles, with their 'expected' limits under permutation, and the resulting histogram for the null distribution of the test statistic. Though these two plots are not of the same type (unlike the previous example of multiple histograms) it is natural to hold these related graphs together in a single construction, the multi-plot *MultiPlot2*. Another similar example is seen in this Bristol Channel ws workspace (*MultiPlot3*), of the dendrogram (*Graph6*) under flexible beta clustering, together with its associated line plot of the cophenetic correlation against the range of beta values (*Graph7)*. PRIMER 7 now automatically packages such naturally related plot windows into a single multi-plot construction, essentially to neaten and simplify the 'house-keeping' of the Explorer tree rather than as a primary presentational tool – a multi-plot is often best thought of as a collection of thumb-nail graphs, each of which can (and should) always be viewed and manipulated individually by clicking anywhere over the space they occupy in the multi-plot window. In the Explorer tree the individual plot names are therefore all listed under the multi-plot name, e.g. *Graph 4* and *Graph5* under *Multiplot2*, and *Graph6* and *Graph7* under *Multiplot3*, etc, and it is often convenient to roll-up the individual plots by clicking on the rolled-out icon ⊟ in front of the multi-plot name, which is then replaced by ⊞, the rolled-up icon, with the individual plot names now hidden (but not, of course, deleted). This is particularly useful when large numbers of component plots are automatically created, as in the **Histogram Plot** on large numbers of variables, or in the next section, the new availability in PRIMER 7 of MDS plots in higher numbers of dimensions, with a run of **Analyse>MDS>Non-metric MDS (nMDS)>** (Min. dimension: 2) & (Max. dimension: 10) generating 9 ordination plots with their 9 associated Shepard diagrams, all automatically combined into an 18-component multi-plot (or 19 plots if the option is also taken to show the *scree plot* of stress *vs*. dimensionality). Of course, configurations in more than 3-dimensions can only be displayed by showing 3 axes at a time, but viewing the change in Shepard diagrams as dimensionality increases, in a single multi-plot, can be instructive.

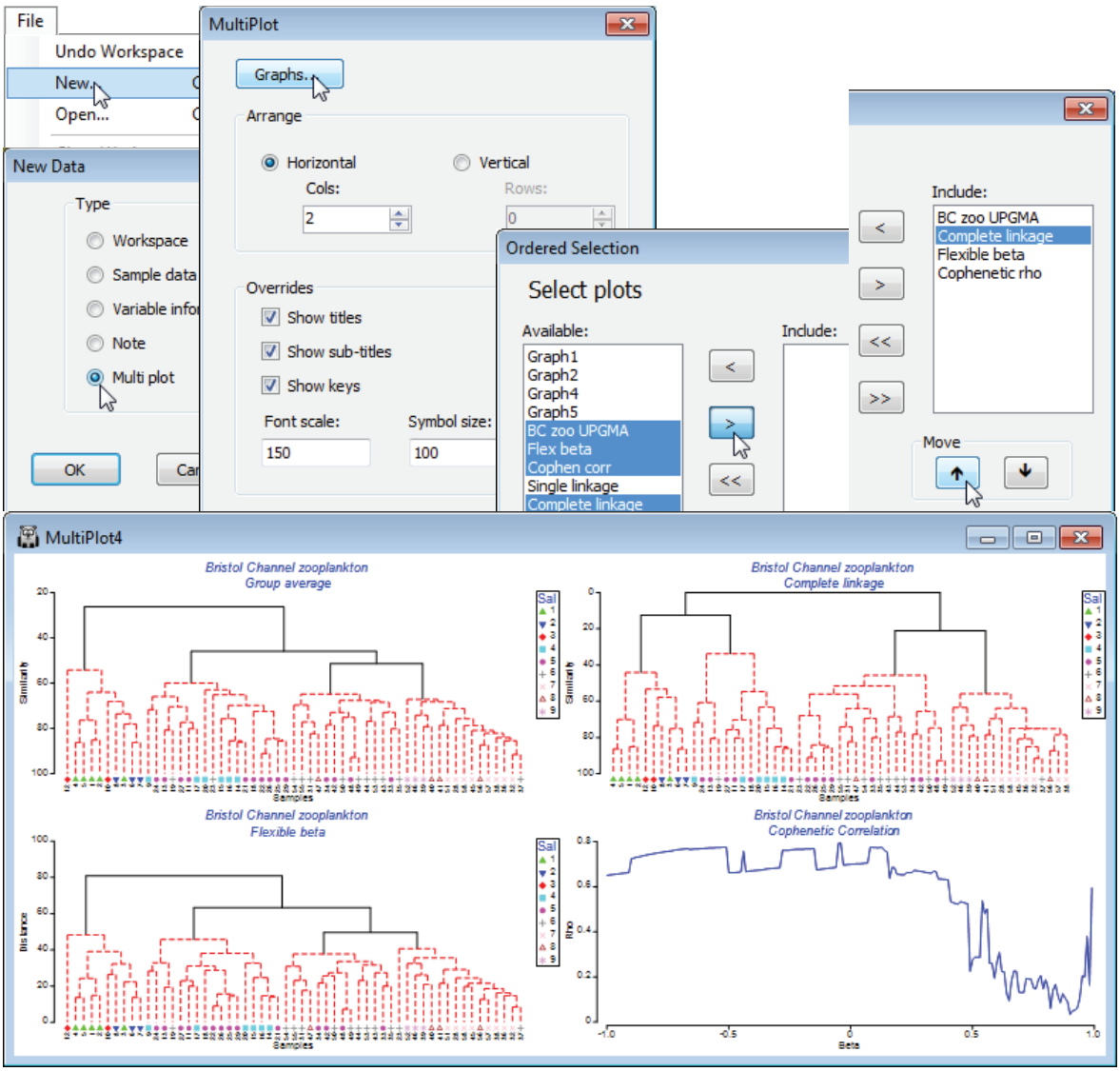
User creation/ manipulation of multi-plots

v7

In addition to automatic generation by certain routines, multi-plots can be created and populated by the user, to hold sets of related plots in a single composite window. **File>New>**(•Multi plot) gives a dialog box headed by a **Graphs** button which allows an ordered selection from the windows for most of the single plot types in the workspace (for a contiguous set from the Available list, click on the first and last name whilst holding down the Shift key, or for a non-contiguous set, on individual names whilst holding down the Ctrl key, in standard Windows fashion). Then move them to the Include list with > and re-order them in the required sequence with the ↓ and ↑ arrows. This sequence is followed either in horizontal rows or vertical columns with the choice of arrangement being chosen in the Multiplot dialog box. Given that the component plots can be of different types, the options for global change of features applying to all plots is inevitably limited. However, this dialog box – which can also be returned to from a completed multi-plot using **Graph>Special** – does allow for global setting of overall font and symbol sizes (where not limited to fit sample axes, as in the example below) and global suppression of main titles, sub-titles and keys. Clicking on individual plots in the multi-plot returns the active window to that specific graph, allowing the usual full range of changes to be made to all its properties. When this window is closed ( ), the multi-plot is again the active window and incorporates those changes (unless globally overridden).

v7

The workspace Bristol Channel ws from C:\Examples v7\BC zooplankton should still be open, and it would be advisable to change the various dendrogram plot windows to more recognisable names, reflecting the different linkage methods, e.g. *Complete linkage* or *Flexible beta* (the linkage method can be found in the results window above each plot). Create a multi-plot with **File>New>**(•Multi plot)>(Arrange•Horizontal>Cols: 2) & (Font scale: 150) & **Graphs**, including the group average (UPGMA), complete and flexible beta dendrograms and the latter's cophenetic plot, in that order. The plots will be placed in a 2-column stack (thus of 2 rows), with the order reading horizontally.



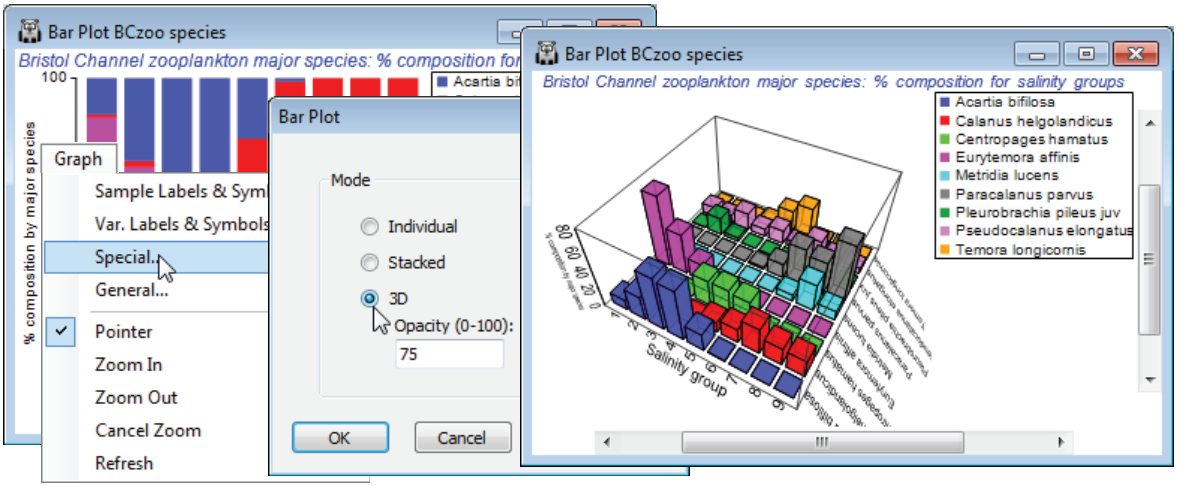
Plots menu

v7

v7

PRIMER 7 has a new **Plots** main menu, available when the active window is a data matrix. This brings together a number of standard plot formats, and more specialised ones, relevant to a range of multivariate analyses. Some are similar or identical to those in PRIMER 6, e.g. **Plots>Dominance Plot**, **Geometric Class Plot** and **Species-Accum Plot** (Section 16), and **Draftsman Plot** (already seen in Section 4 and again in Section 13). Others are new to PRIMER 7: the **Histogram Plot** was seen earlier (Section 4), as were **Scatter Plot** (Section 5) and **Surface Plot** (end of Section 4 and in Section 5). A significant new feature in PRIMER 7, the specialised **Shade Plot**, was introduced in Section 4 as a means to aid pre-treatment choice and is discussed in more detail in Section 10, in the context of interpreting species patterns across samples. There is a similar context for **Line Plot**, in the form of a multi-plot created by **Wizards>Coherence plots**, seen in Section 10, and the other three plot types in this menu are also standard graphic constructions: a **Bar Plot** can show relative composition of different species in each of a set of (averaged) samples, and **Box Plot** and **Means Plot** (Section 15) provide standard univariate tools pre- and post-hypothesis testing for a one-way layout, e.g. the latter giving confidence intervals for *effect sizes* (generalised to multivariate data by the *bootstrap region plots* of Section 17, under **Analyse>Bootstrap Averages** on resemblances).

The **Plots** menu items are mainly exemplified in the analysis sections in which they are most likely to be useful, but three of the standard graphic displays are given below for the Bristol Channel data.



Save and close the Bristol Channel workspace, Bristol Channel ws.

Workspace planning

To conclude this section, it is worth remarking that care taken in structuring workspaces will often pay dividends if the analysis results need to be returned to later. An Explorer tree represents a single workspace. It can contain several starting data matrices, the properties (factors etc) of each being accessible to the others if they are based on an overlapping set of identical sample labels. (As seen previously, factors or indicators associated with a particular data matrix are automatically available to other sheets on the same branch, and on different branches by using the **Factors> Import** button). Such is the power and reach of PRIMER to quickly generate many plot and results windows that the user will probably find it a constant battle to keep workspaces down to a manageable size, both from the viewpoint of ease of navigation around them and of the size of a workspace file that needs to be transmitted to others (we have repeatedly seen the convenience of saving all the information in the workspace in a single file with a **File>Save Workspace As** step).

Firstly, it makes obvious sense to keep analyses on different studies in different workspaces but, also, analyses on different components of a single study may sometimes be better carried out in separate workspaces. The key criterion for using a single workspace is whether two data sets are to be combined or input to an analysis step together (e.g. species and matching environmental arrays), even if this is just as component parts of a single multi-plot. If not, they are probably best held and saved as separate workspaces. Multiple launches of the PRIMER desktop are straightforward, each with a different workspace (if the same workspace is opened twice – which is perfectly possible though should usually be avoided because of the likelihood of confusion when saving! – the second will be a copy of the saved version of the first workspace). Parallel desktops will not interfere with each other, and are never linked in any way. The only means of transfer between them is by saving individual sheets (e.g. data as *.pri) and then opening (a copy) of that file into the other workspace.

Secondly, 'housekeeping' within a workspace is important for intelligibility: key sheets of data, resemblances, results or plots – anything that needs to be selected as the active sheet for a further analysis, or a result or plot window that has been copied and saved to an outside presentation or manuscript – should be renamed in a meaningful way (all names in the Explorer tree need to differ, though PRIMER will ensure this by adding *(2), (3), ..* to the end of any name you supply which is already used in the workspace). Also, it is usually advisable to delete clutter, e.g. analyses you now realise were flawed or sub-optimal, using **File>Delete …** or **Delete Item** on the right-click menu when the window to be deleted is highlighted (branch entries below that will be deleted too). Use **File>Undo Workspace** if you make a mistake and want the excised portion back again! If you decide the sub-optimal analysis needs to stay in, as a reminder, then roll-up that particular branch (with ⊞) and attach a note to the window above the ⊞ icon, with right-click **Add Note**. In fact, it is desirable to make good use of this annotation feature more generally, to aid navigation.

Finally, some studies are sufficiently extensive, with data accreting over time, that it is advisable to resave the workspace with a modified name from time to time, so that at least an earlier version can be returned to should a disaster happen to the current workspace! Or it might be efficient to save the current data matrix (or matrices) in PRIMER binary format, *.pri, thus retaining all existing factors and indicators, and re-open this in a clear workspace for the next phase of the analysis.

8. Multi-dimensional scaling (*Non-metric nMDS*, *Metric mMDS, Combined MDS*)

Rationale for *n*MDS & *m*MDS

v7

Chapter 5 of the CiMC methods manual describes the operation and rationale of multi-dimensional scaling (MDS) ordination, **Analyse>MDS**. The aim of MDS is to represent the samples as points in low-d space (often 2-d or 3-d, but PRIMER 7 will now compute MDS solutions for any specified range of dimensions), such that the distances apart of all points are as closely matched as possible to the relative dissimilarities (or distances) among the samples, as measured by the resemblance matrix calculated on the (pre-treated) data sheet. The definition of 'closely matched' for the most commonly used form of MDS, **Non-metric MDS (nMDS)**, is that the rank order of dissimilarities among pairs of samples are preserved in ranks of the corresponding pairwise distances in the final ordination plot. The interpretation of a (successful) *n*MDS is therefore straightforward: the closer points are to each other the more similar is their community composition (or suite of environmental data, biomarker responses, particle size distributions, or whatever the variables represent).

v7

PRIMER 7 also provides the more parametric technique of **Metric MDS (mMDS)**, which seeks to interpret the entries in the resemblance matrix as actual distances, so that samples with distance/dissimilarity *d* are placed at distance *d* in the ordination plot. The key distinction is that the Shepard diagram (a scatter plot of resemblances, *x,* against ordination distances, *y*) is fitted by a straight line in *m*MDS but by a general (non-linear) increasing function in *n*MDS. The much greater flexibility of *n*MDS makes it more suitable for displaying typical community data in low dimensional space, but low-d *m*MDS plots have a useful role to play in ordinations on very few points (as for some means plots) and in region estimates for means (Section 17), or where the resemblance coefficient is, or behaves very like, a genuine distance. Examples might be for normalised Euclidean measures on environmental-type data or community data with low sampling variability from a *short baseline of change* (i.e. relatively little species turnover). Data with more typical sampling variability, but still over a short baseline, can sometimes be well represented in low-d by *threshold metric MDS* (*tm*MDS, Fig. 5.12, CiMC), in which the Shepard plot is fitted by a straight line but not through the origin. Unlike *n*MDS, (*t*)*m*MDS plots thus have a measurement scale interpretable in terms of the resemblances, though all forms of MDS plot are arbitrarily rotatable and reflectable in the axes.

Metric MDS is <u>not</u> Principal Co-ordinate Analysis (PCO), as available in the PERMANOVA+ add-on to PRIMER, though this is a common misconception. PCO is a projection technique from high to low-d, via an eigenvalue decomposition – a generalisation of PCA, see Section 12. The *n*MDS and *m*MDS algorithms are iterative searches, not guaranteed to converge to the optimal solution, hence the need to run them for many random restarts. The default in PRIMER 7 is 50 restarts but if the run time for a single one is not an issue, it is always worth doubling that number, to ensure that a solution is found which is, at least, near-optimal. A working criterion for deciding that enough iterations have been performed is that the same (lowest) stress value is obtained from more than a handful of the restarts. *Stress* measures the scatter in the Shepard plot, i.e. how faithfully the high-d relationships are represented in the low-d ordination – for interpretation of stress values see CiMC.

Combined MDS & 'Fix Collapse'

v7

A further new feature of *n*MDS in PRIMER 7 is the ability to minimise a combination of two stress functions, equally mixed – this has potential application, for example, to combining information on a common set of samples from community matrices (typically using a biological resemblance, such as Bray-Curtis) with that from physical variables (usually requiring Euclidean distance) in a single ordination, a **Combined MDS** plot. A more commonly needed requirement is implemented within the *n*MDS routine: the ability to mix a small amount of a metric solution with the predominantly non-metric one, preserving the flexibility of *n*MDS whilst implementing a (✓Fix Collapse) of the non-metric solution which can occur if a sample, or set of samples, has greater dissimilarity to all others than any dissimilarity within either set. Ranks then carry no information about the relative spacing of the two sets and even a very small amount of metric MDS information is enough to fix this indeterminacy. This will often be a better option than using the **Graph>MDS Subset** routine on a box drawn around the main group of points, excluding the outliers causing the difficulty.

Diagnostic tools for MDS plots

In addition to the ability in a previous PRIMER version to **Graph>Special>Overlays**>(✓Overlay clusters) from a dendrogram onto a related MDS ordination, in order to judge agreement between these differing low-d displays of high-d data, PRIMER 7 now provides a wide range of diagnostic tools to monitor convergence of the ordination and the adequacy of the low-d representation. The

v7 iterative search process can be viewed in real time with (✓Configuration plot)>(✓Animate) and, as with most such animations (including rotating 3-d MDS plots with **Graph>Spin**), recording this, with standard video controls in an *.mp4 file, is now possible. The behaviour of stress over a range of dimensions is seen in a (✓Scree plot) and Shepard plots for all specified dimensions viewed in conjunction, in a *multi-plot* (see the previous section) along with the configurations. Points which the ordination is unable to place well are identifiable from the Shepard diagram by clicking on outliers in this scatter plot. An alternative now available to drawing cluster contours for specified similarity thresholds is a juxtaposition of a 2-d ordination with the full dendrogram in the third dimension, taking **Special>**(**Main>**Plot type•2D>✓3D project) & (**Overlays**>✓Overlay clusters). Also in 2- or 3-d, (**Overlays>**✓Join pairs) simply joins pairs of sample points in the ordination plot which have similarities greater than a specified value, and (✓Overlay minimum spanning tree) will connect ordination points according to the minimum length (branching) path connecting them all, through the dissimilarities in the resemblance matrix (not distances in the low-d ordination). Both methods therefore may allow identification of points in the ordination which do not reflect well the underlying dissimilarities. Another new PRIMER 7 feature here is the ability quickly to match different ordinations of the same set of sample labels (e.g. with different stresses, metric *vs* non-metric etc), by optimal rotation, reflection and scaling of two configurations by Procrustes analysis, taking **Graph>Align Graph** and supplying the configuration plot which it is attempted to match.

Overlaying factors or other data (bubble plot)

The ability to display, on any ordination plot, external structure such as factors (e.g. for sites, times etc) by **Graph>Sample labels & symbols** is a fundamental interpretational tool, as is the addition of time (or unidirectional space) trajectories, via **Graph>Special>Overlays>**✓Overlay trajectory, and supplying a numeric factor which determines the order in which points are joined. This, too, is greatly expanded in PRIMER 7 by allowing multiple trajectories, e.g. a common time course drawn in separate colours or line styles for different sites, using (✓Split trajectory) and then supplying the site factor. Also, there is the capacity to view the evolution of a trajectory (or multiple trajectories) dynamically, in an animation which can be recorded (as noted above) in an *.mp4 file.

v7

Bubble plots (**Graph>Special>Main**>✓Bubble plot) overlay circles whose sizes reflect values of one of the variables (e.g. species) used in constructing the MDS, or of an external variable such as environmental information. Again, bubble plots have been greatly extended in PRIMER 7, from improvements in flexibility, such as the degree of user control over definition of the bubble key (**Key**>Key values), to several new features: drawing simple bubble plots with a user-defined image (**Key>**✓Use image) displayed at different sizes; the automatic availability of '3-d effect' bubbles on 3-d ordinations (justifying the term bubble plot rather than circle plot!); the display of bubbles for single variables in different colours dependent on the levels of a factor, by **Key>**(✓Use symbol colours) when **Sample labels & symbols**>(Symbols✓Plot)>(✓By factor) is selected; and, perhaps most significantly, *segmented bubble plots* are introduced. These display several variables on the same MDS plot as different-sized segments of a circle or sphere, in differing colours and segment positions for the differing variables (under ✓Bubble plot>**Change**, add more variables to Include).

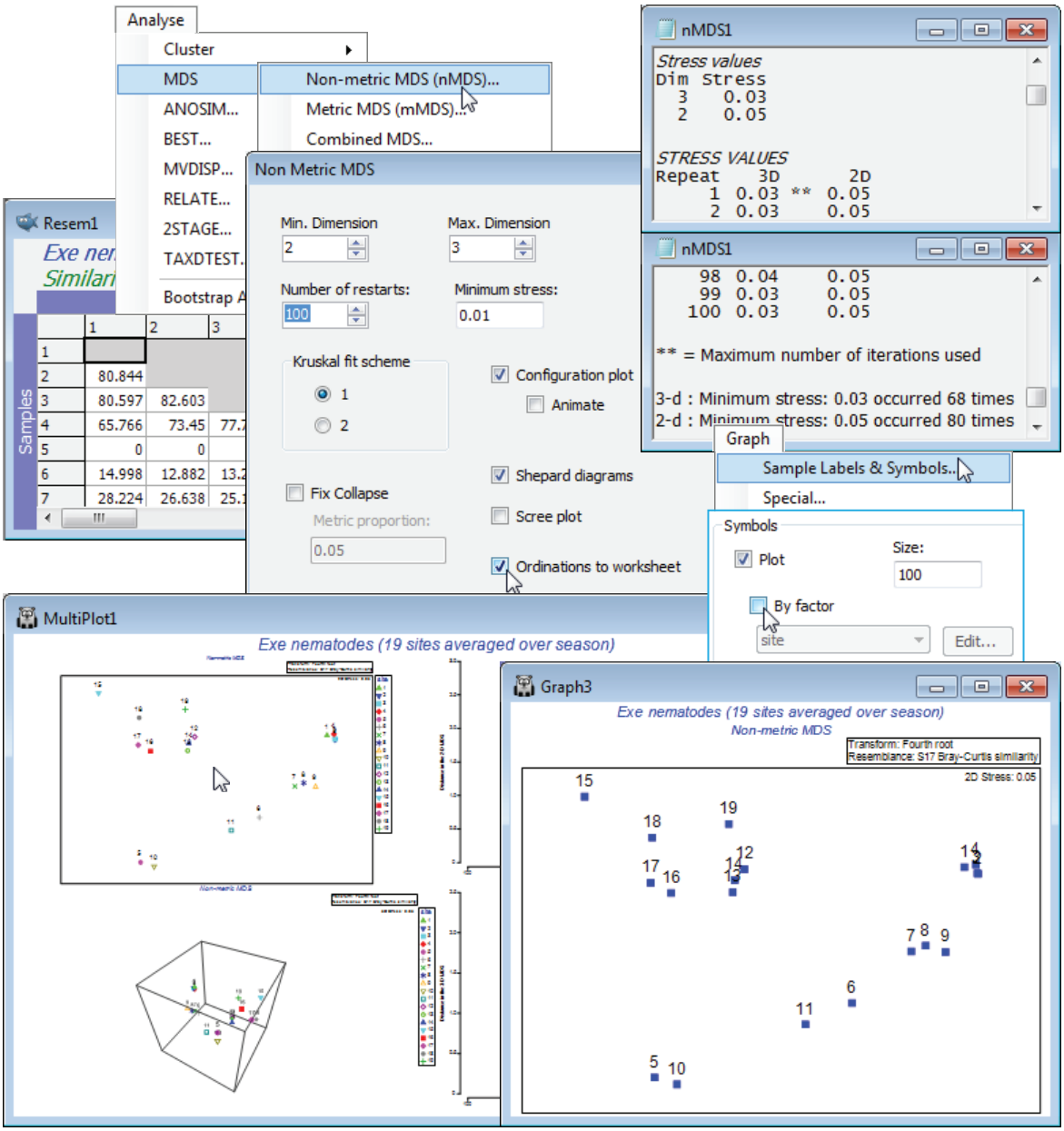
v7

Running an *n*MDS (Exe nematodes)

From the directory C:\Examples v7\Exe nematodes, **File>Open** the workspace Exe ws, last seen in Section 6, of the sediment nematode communities at 19 inter-tidal sites in the Exe estuary. If this does not exist, open the data file Exe nematode abundance(.pri) in a clear workspace, and re-run the UPGMA clustering, with **Pre-treatment>Transform (overall)>**Transformation: Fourth root and **Analyse>Resemblance>**(Measure•Bray-Curtis similarity)&(Analyse between•Samples), then **Analyse>Cluster>CLUSTER>**(Cluster mode•Group average), and on the resulting dendrogram, **Graph>Special>**(Slicing✓Show slice)>(Resemblance: 30)>**Create factor**>(Add factor named: 30% slice). The Bray-Curtis similarity matrix is in Resem1 and the dendrogram Graph1.

With Resem1 as the active window, take **Analyse>MDS>Non-metric MDS (nMDS)** and options of (Min. Dimension: 2)&(Max. Dimension: 3)&(Number of restarts: 100)&(Minimum stress: 0.01) &(Kruskal fit scheme•1)&(✓Configuration plot)&(✓Shepard diagrams)&(✓Ordinations to work sheet) leaving the other boxes unticked for now. The outcome is a results window, nMDS1, and a multi-plot MultiPlot1 which, if unrolled in the Explorer tree (either by clicking the ⊞ in the tree or by clicking on any of the plot components in the multi-plot), shows four plot windows, probably named Graph3 to Graph6. For the first one, remove the spread of symbols with **Graph>Sample labels & symbols**, unchecking (Symbols>By factor), to leave just the labelled default symbols.

v7



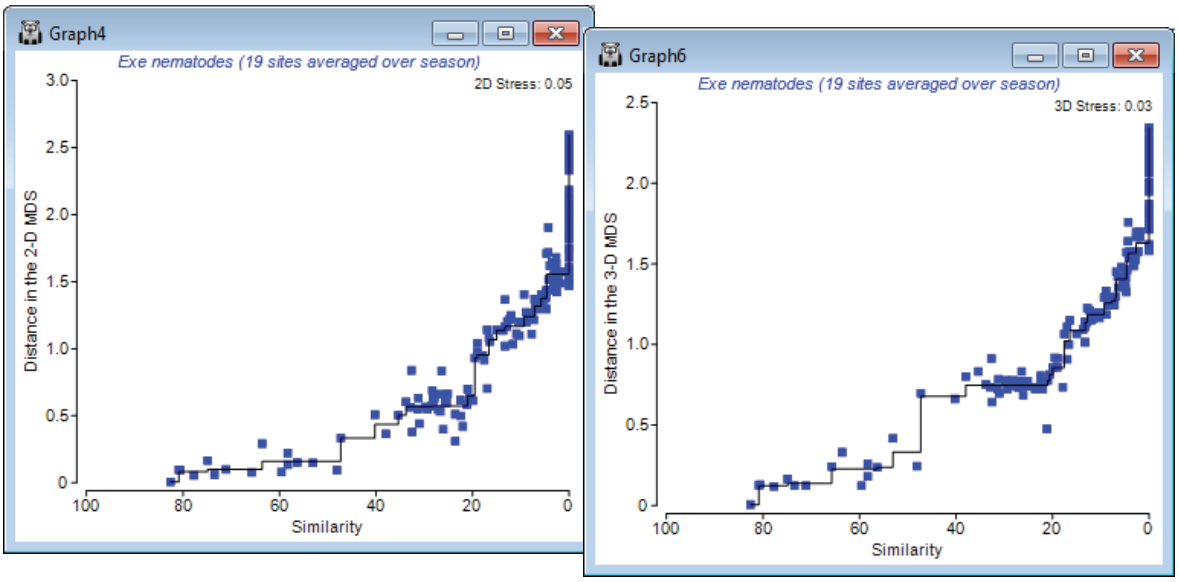
MDS results window

As with all results windows, the MDS1 window first lists the resemblance sheet on which analysis is carried out, and whether it was under any selection on the original sample set (given as ranges of sample numbers) but the main function of this window is to report the stress the iteration converges to, from each of the 100 restarts (independently for each dimensionality, here just 2-d and 3-d). The same lowest stress values of 0.03 for the 3-d solutions, and 0.05 for 2-d (stress must be higher in lower dimensions), are found over 50% of the time, so is very unlikely to be bettered by any further restarts. Note that your run of MDS will not produce exactly this sequence of stress values, because restart layouts are randomly drawn, and the random seed is taken from the computer's clock. Each restart involves an iterative cycle (alternately fitting monotonic regression to the Shepard diagram and adjusting the configuration points by steepest descent – see Chapter 5 of CiMC). If the process has not converged within 100 cycles then it terminates: this happens here for a few of the restarts with the 3-d solution, indicated by ** after the quoted stress. (It probably suggests that two equally good solutions are available, and the algorithm cycles back and forth between them, which happens often when the structure can fit easily into low-d, so that stress is very low).

Shepard diagrams

That the stress is low here is also evident from the Shepard diagrams for the 2-d and 3-d solutions (possibly Graph4 & Graph6). They are scatter plots of distances between samples in the ordination against original (dis)similarities, in which the deviations of the points (blue) from the from the best-fitting monotonic increasing regression line (black) are seen to be very low. When all points lie on the line, all statements of the following form are true: 'site 5 is closer to site 10 in the ordination

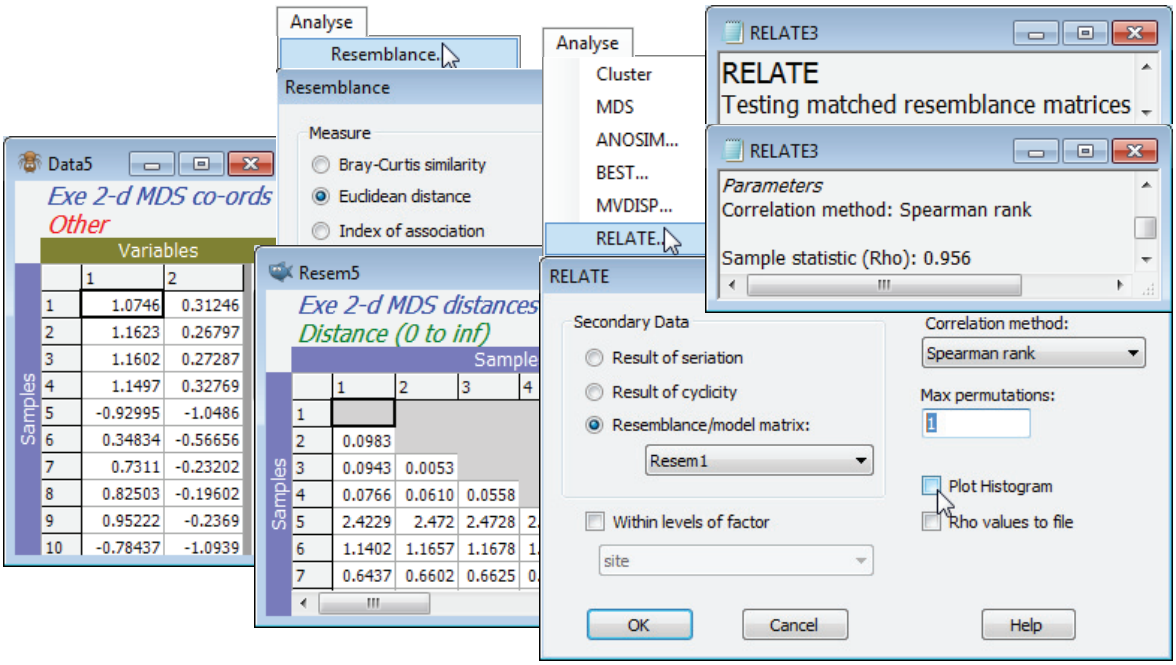
than site 14 is to site 6 because the dissimilarity between sites 5 and 10 is smaller than that between 14 and 6', i.e. rank order relationships are exactly preserved and stress is zero. Note the regression does not need to be linear (through the origin) as in *m*MDS, and it is certainly not so here: sites 14 and 6 are twice as dissimilar (~ 90%) as sites 5 and 10 are (~45%) but they are not placed twice as far apart in the plot (as *m*MDS would try to force). This capacity to shrink/stretch the dissimilarity scale in conversion to a distance is what gives *n*MDS its flexibility.



Dissimilarity preservation as a matrix correlation

v7

One can also ask how well the (Euclidean) distances among points in the *n*MDS plot correlate with the dissimilarities in the resemblance matrix. The former are calculated by running the ordination co-ordinates (output to Data4 and Data5 by the ✓Ordinations to worksheet instruction in the above example) through **Analyse>Resemblance>**(•Euclidean distance). Then, just as for the Cophenetic correlation heading in the Section 6 cluster analyses, which was carried out on the same Exe data, a matrix correlation between these two triangular matrices requires a run of the **Analyse>RELATE** routine (Section 14), e.g. with the distance matrix as the active sheet and the dissimilarities Resem1 as the secondary data (or vice-versa). The only difference this time is that the option to compute a rank correlation such as Spearman should be taken (a *rank Mantel*-type correlation), since this is *n*MDS and the Shepard plot is not linear. (It is often overlooked that Pearson correlation measures only linearity of a relationship – a stress of zero corresponds to Spearman $\rho_S = 1$ but Pearson $\rho < 1$, when the increasing relationship is perfect but not linear). The permutation test in **RELATE** is not required since $\rho = 0$ is not a sensible null hypothesis, so set Max permutations: 1 and uncheck the Plot Histogram box, giving $\rho_S = 0.956$ for the 2-d *n*MDS and 0.965 for the 3-d configuration.

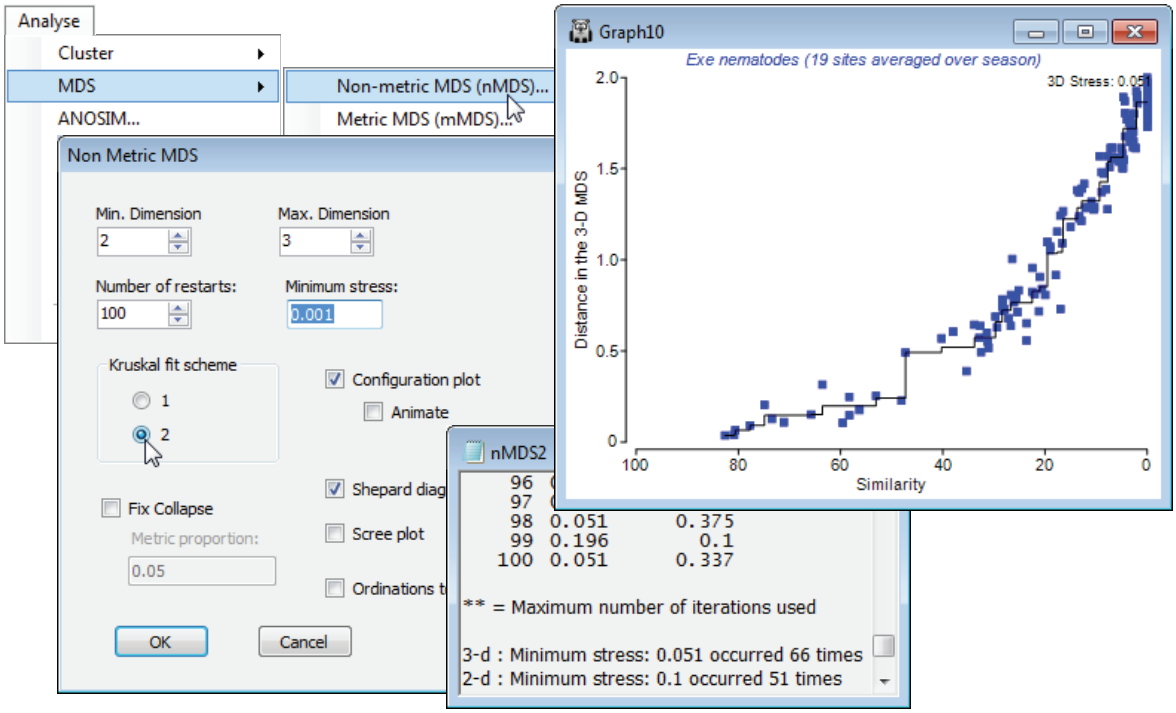


Accuracy & fit scheme

For the MDS run above, two of the defaults taken for options in the MDS dialog were (Minimum stress: 0.01) and (Kruskal fit scheme•1). Changing the former from 0.01 to 0.001 would decrease the lower threshold of stress at which the iteration decides that it has effectively reached a perfect solution but, more usefully, also increases the accuracy with which stress values are reported in the results window. Reporting stress to a third decimal place can be useful in deciding whether a batch of restarts with the same stress, to two decimal places, are really the same solution. However, it is unwise to take small differences in stress too seriously: solutions with nearly the same stress will usually lead to the same interpretation. A low-dimensional ordination is only an approximation to the real high-dimensional pattern, in any case, and not necessarily a very good one. (This is the reason that most of the substantive analysis, like hypothesis testing, takes place on the resemblance matrix and not in a low-d ordination space. It therefore misses the point to worry unduly about whether a low-d plot is the optimum placement of the points or one that is very nearly optimal – both are only approximations to the truth). In fact, it can be quite revealing to look at repeat MDS runs with only one restart, which much of the time will therefore converge to an inferior solution, and observe which points differ from their placement in the optimal solution.

The (Kruskal fit scheme•1) option is by far the commonest choice for practical *n*MDS. Essentially, it allows dissimilarities which are equal (tied ranks) to be represented in the final ordination by distances which are not equal, whereas (Kruskal fit scheme•2) constrains those plot distances to be equal. The latter can be an unhelpful constraint in any situation in which there is a complete turn-over of species across some samples. For example, along a strong environmental gradient, such as water depth say, there could already be complete species turnover in benthic organisms between sedimentary sites at 5m and 100m (dissimilarity = 100%), but two sites at 2m and 200m, or at 1m and 500m, cannot give a larger dissimilarity than this. If 100% dissimilarity is to be represented by exactly the same distance in the ordination of samples widely spread along this depth gradient, it is inevitable that an arched (or to be more precise, staple-shaped) solution will result from what is actually a strong linear gradient. This is one of several explanations for the *arch effect* seen in other ordination techniques (such as PCO), to which *n*MDS is less prone because of the flexibility under (Kruskal fit scheme•1) to represent the set of 100% dissimilarities by different plot distances.

The above Shepard plots for the Exe nematode data – which is an example of large-scale species turnover – show this flexibility, in representing the many similarities of zero (dissimilarities of 100) by distances from about 1.6 to 2.4 in the 3-d plot, with stress of 0.033 (when run to 0.001 accuracy). Re-running for (Kruskal fit scheme•2) is seen below to force these distances closer to equality (1.8 to 2) and slightly degrades the solution (stress = 0.051 for the 3-d plot). The extra d.p. in quoting stress has not here demonstrated much of a spread of near-optimal solutions at a finer scale: the lowest stress of 0.051 is obtained at only a little lower frequency than before (66 of 100).



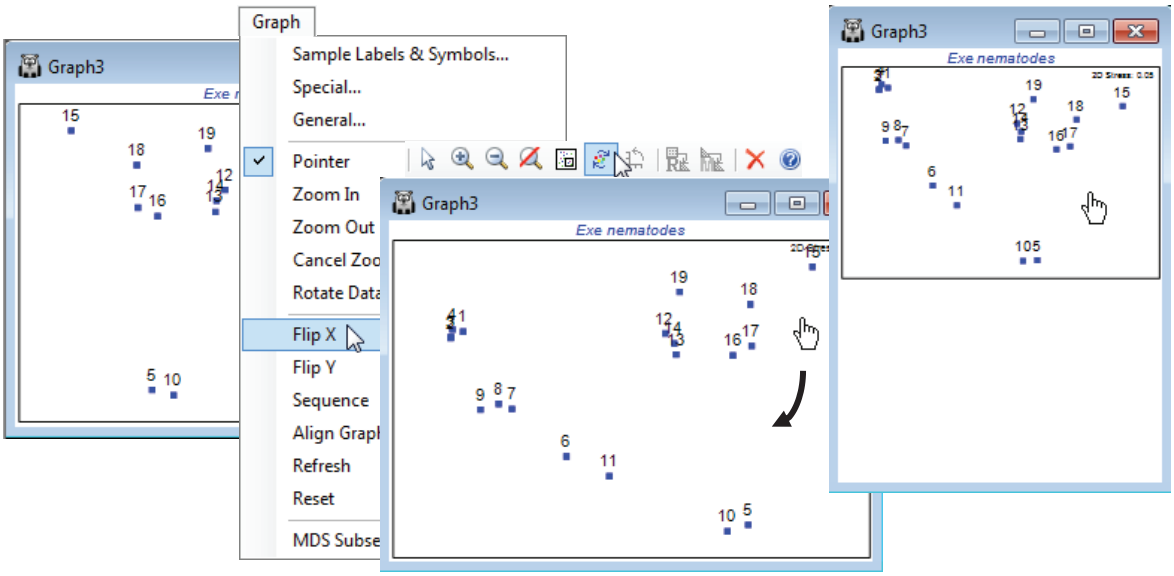
Graph menu: rotating and flipping the 2-d ordination

*n*MDS ordinations (unlike *m*MDS) have no meaningful axis scales for the configuration, since they use only ranks rather than original dissimilarities. There are also no defined directions for plot axes (unlike PCA, Section 12, or PCO, PERMANOVA+ add-on), so MDS plots (of any type) can be arbitrarily rotated, with **Graph>Rotate Data**, or by clicking on the Rotate data icon on the tool bar. The cursor changes to a hand and by clicking, holding and dragging, the plot can be rotated continuously within its current rectangular frame. Purely for reasons of presentation, wide, fat plots are usually preferred to high, thin ones, so the MDS algorithm defaults to orienting the plot with its major axis across the page (which it does by inputting the final MDS co-ordinate positions into a PCA, and the internal axes scales are then determined by the major axis being normalised to mean zero and variance one – do not confuse this purely presentation feature with running a PCA on the original data!). However, any orientation is equally meaningful so if a specific orientation is needed for external reasons (e.g. to attempt to match a community ordination to the geographic location of the samples), once the desired rotation is achieved the cursor can be changed back to by **Graph>Pointer**. A new addition in PRIMER 7 is the **Graph>Reset** option, which allows you to reset the plot to the default orientation of the original MDS run, when the co-ordinate positions of the points on the plot will match those in the worksheet which may have been requested by (✓Ordinations to worksheet). It was noted in connection with cluster dendrogram rotations that the current ordination co-ordinates, after rotation, can always be retrieved by **File>Save Graph Values As**, to obtain an external text file (which could always be read back into PRIMER on the odd occasion when this might be required – though because of the arbitrariness of MDS axes it is generally not desirable to use single axis co-ordinates in subsequent regression/correlations, e.g. linking to abiotic variables).

v7

In a similar way, MDS plots can be reflected on axes by **Graph>Flip X** or **Flip Y**, and the default configuration returned to by **Graph>Reset**. Though it is clear that much information about the axes (scaling, orientation, reflection) is arbitrary with *n*MDS, what is not arbitrary of course – in fact central to the method – is relative distances apart of points. Changing the aspect ratio of the points in an MDS plot (shrinking or expanding it along one axis only) destroys the key inferences, of the form 'samples 15 and 16 are closer together than 5 and 6 therefore they are more compositionally similar' (even though the direction of 15 to 16 is perpendicular to that of 5 to 6). Within PRIMER, this is not a concern. For the MDS plot, Graph3, in the Exe workspace, try flipping and re-orienting the plot and reshaping the window and you will see that the plot preserves the original relationships among the points rather than, for example, taking up the shape of the new window (as it would do with a cluster dendrogram). You should, however, be careful not to reshape the plot later, having saved or copied it via the clipboard to Powerpoint or other graphics presentation package.

v7



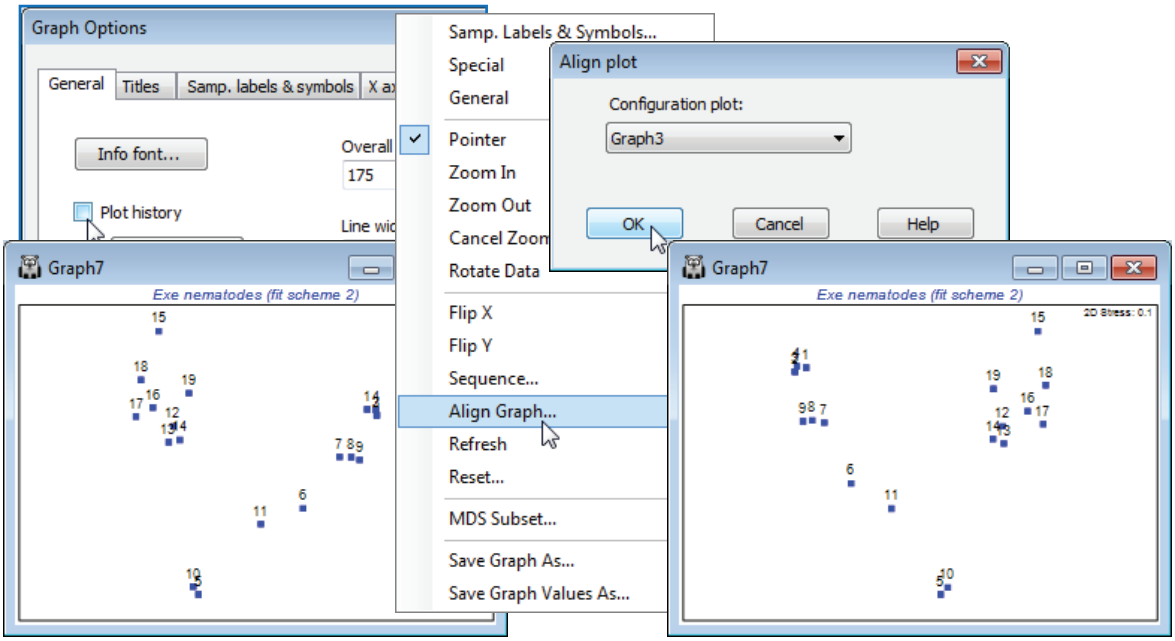
Align graphs automatically

A new feature in PRIMER 7 is the ability to automatically reflect in the axes and rotate an MDS configuration (and rescale it, if necessary) such that it best matches the pattern of another supplied configuration. This can make it quicker and simpler to compare patterns from several different ordinations of the same sample labels, trying out the effect of different pre-treatments, resemblance measures or simply re-runs with more restarts etc. This is accomplished by Procrustes analysis (see the CiMC reference list for the Gower 1971 reference) and operates only on two graphs at a time.

v7

In the current Exe nematodes workspace Exe ws you should already have at least two 2-d *n*MDS plots for the 19 sites (e.g. Graph3 and Graph7 under Kruskal fit schemes 1 and 2). Automatically rotate and flip the active window (say Graph7) to match that of Graph3 by right-clicking over the plot (or taking **Graph**) to get **Align Graph>**(Configuration plot: Graph3). Note that the examples above and below have used the **General** and **Title** tabs on the Graph Options dialog to remove the history box (uncheck Plot History), amend main title, remove subtitle, change overall font scales etc. All these features and the **Samp. labels & symbols** tab are exactly as covered in Section 6.

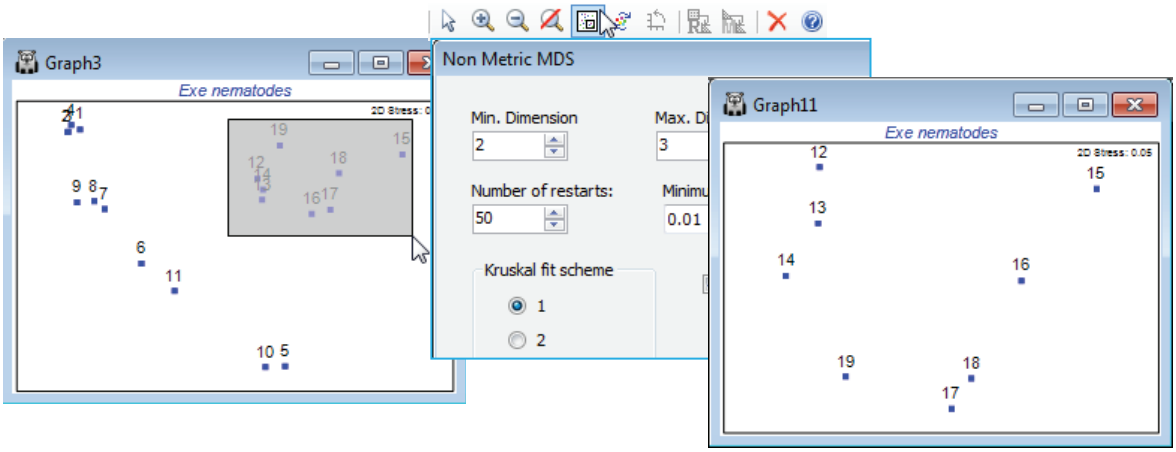
v7



Zoom & MDS subset plots

v7

It is scarcely necessary here, with only 19 samples, to zoom in on part of the plot in order to see the detailed structure, though this is possible using the **Zoom In**, **Zoom Out** and **Cancel Zoom** options on the **Graph** menu (also accessible from the , and icons on the Tool Bar), as described for zooming in on dendrograms in Section 6, and this may be useful for very cluttered MDS plots, with rather too many points. However, it is not the only, or even the best, possibility. If attention is to be focussed on only a subset of points, then it is desirable for the 2-d ordination to display those points in the best possible relationship to each other, matching their (smaller set of) dissimilarities. This can be accomplished more accurately, since it is no longer necessary to display their relations with all the omitted points. In other words, the MDS should be repeated from scratch on the subset, and one way of carrying this out is to draw a box on the MDS (with cursor as pointer), clicking and dragging a rectangle around the required subset of points, then taking **Graph>MDS Subset** (or its icon). This automatically selects the surrounded points from the resemblance matrix and displays the MDS dialog box (normally obtained by **Analyse>MDS>…**). For the Exe nematode data, even though the 2-d stress for the full set of samples in Graph3 is low (at 0.05), a re-run on sites 12-19 gives a more accurate display of the site 12-14 relationships (a SIMPROF test, as seen in Section 6, actually divides site 14 from 12 & 13, not something that would be expected from the original Graph3 display, in which 14 appears to sit between 12 and 13).



Special menu for ordination

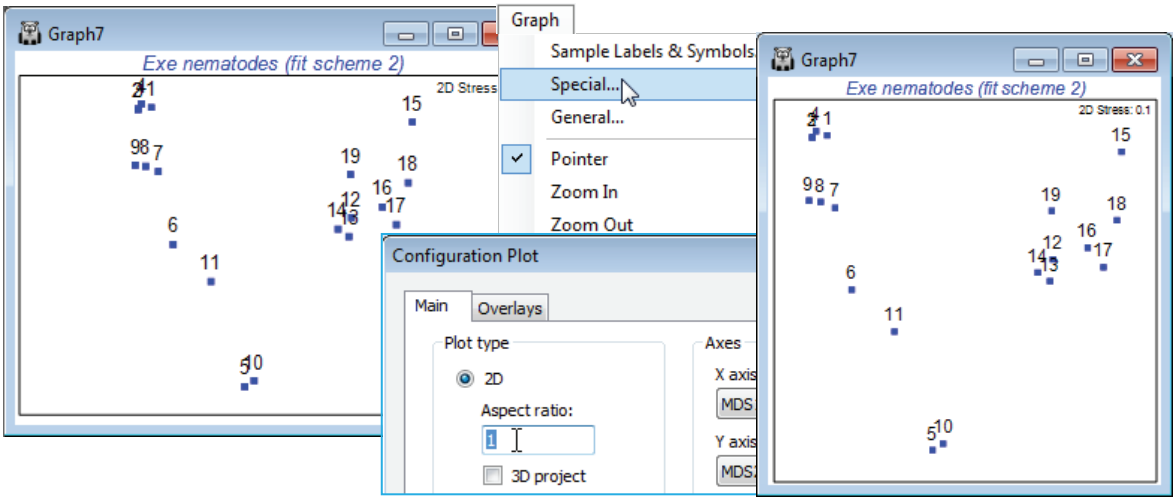
We have seen before that, whilst the **Graph>Sample Labels & Symbols** and **General** menus are universal, having dialog boxes of the same form for all the graphics routines, the **Graph>Special** dialogs are specific to each routine, and this option is at its most extensive for ordination plots from *n*MDS, *m*MDS, PCA (and PCO in the PERMANOVA+ add-on). On the **Main** tab, options include specification of plot types, which axes to plot, setting up of a range of bubble plots, and animated displays of ordination points in time (or other) sequence. On an **Overlays** tab, there are options to overlay temporal or spatial trajectories through the points and vector plots, and various diagnostic aids, such as superimposing cluster results, minimum spanning tree and joining similar samples.

v7

Aspect ratio of boundary

Before moving on to diagnostics it is natural here – having discussed zooming of rectangular boxes within an ordination – to cover the first option on **Graph>Special>Main**, the ability to change the aspect ratio of the boundary of an MDS plot. It was noted above that the aspect ratio of the points in an MDS must never change because this destroys their relative distances apart (in any direction), the key information that a (scale-less) *n*MDS carries. By default, the MDS points are placed within a rectangular border with an aspect ratio of 1.5:1, width to height. This is purely because the default configuration has been rotated to principal axes, as noted earlier, for presentational reasons: most plots are conveniently displayed in landscape rather than portrait format. However, if a different aspect ratio for the border is required (and of course there is no possibility of obtaining this by re-sizing the window displaying the plot!) then, for example, **Main**>Plot type•2D>(Aspect ratio: 1) will produce a square boundary. (Indeed, some practitioners prefer a square boundary for all MDS plots, since axis direction is arbitrary, and early versions of PRIMER did have this constraint.)

v7



Diagnostics for MDS: join pairs

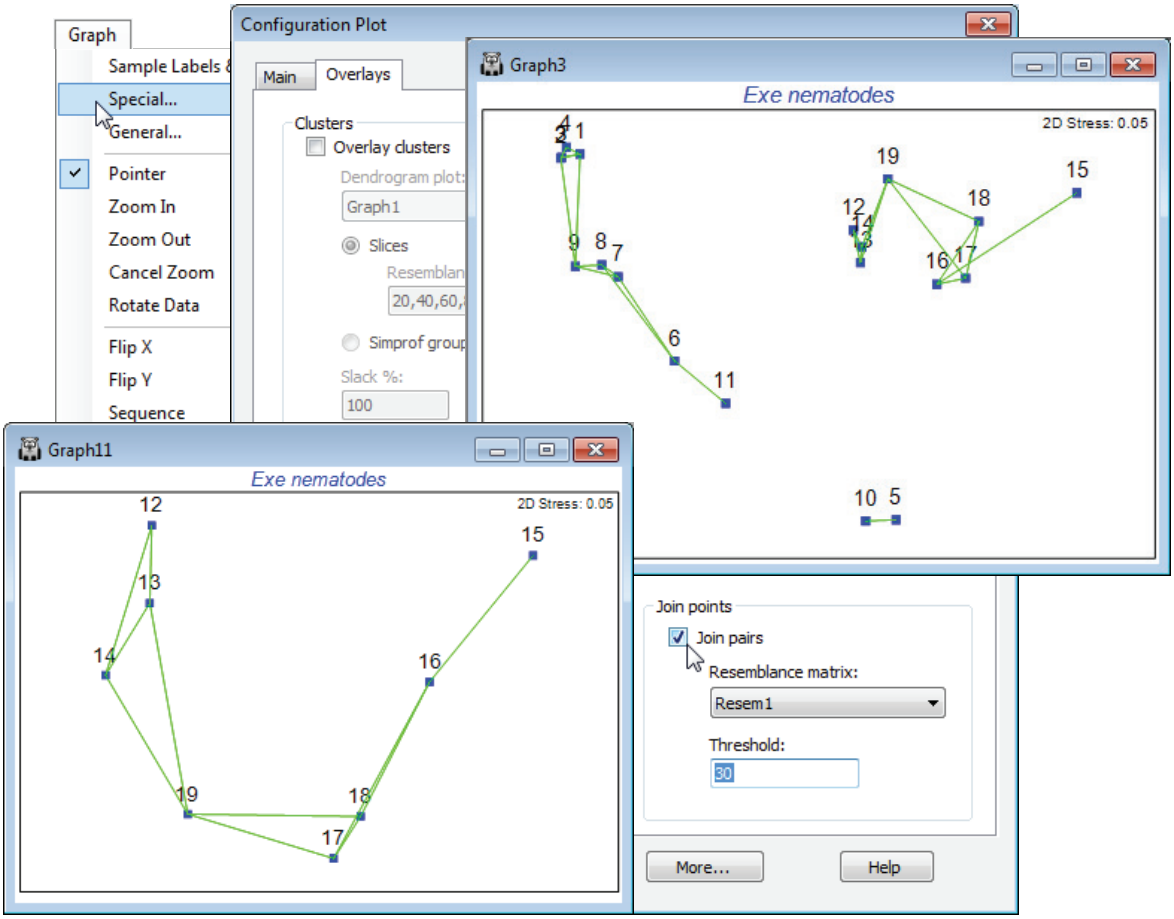
There are a number of new diagnostic tools available in PRIMER 7 for assessing how well a low-d ordination represents the structure in the resemblance matrix. One of the simplest is to join pairs of points on the ordination which have similarity greater than some supplied threshold value. Choice of which value(s) to use requires a certain degree of experimentation, perhaps guided by a cluster dendrogram – too large a threshold and there are not enough joined pairs for an informative plot, too small and the plot is over-cluttered. For the Exe ws data, and the 2-d MDS plot Graph3, take the Overlays tab on the Special menu, i.e. **Graph>Special>Overlays**>(Join points✓Join pairs)> (Resemblance matrix: Resem1) & (Threshold: 30), which joins all pairs with similarity >30%. Whilst there is little conflict with the representation of such dissimilarities (<70%) by distances in the MDS plot, a small amount of inaccuracy (i.e. stress – low, at 0.05, but not entirely negligible) is evident in the way site 15 is connected with 16 but not sites 17 or 18, to which it appears closer in the 2-d plot. Note that if the same Join pairs operation is taken on the subset MDS of the last page, for sites 12-19 alone, this inaccuracy is resolved, as is the slight conflict noted previously for sites 12-14 – the reason site 19 is joined to 14 but not 12 in the full plot is evident from the subset MDS.

v7

Features that carry over to 3-d ordination

It is worth saying, as we start to explore the wide range of tools which can be applied to an MDS or other ordination that, although demonstrated for 2-d plots, most of the techniques are also available for 3-d configurations. This includes all joining operations, such as join pairs and the minimum spanning tree, trajectory and vector overlays, and the new sequence animations. Also bubble plots are extended to 3-d in PRIMER 7. In fact the only Special menu feature which is not so extended is the drawing of cluster contours, but plotting symbols for cluster groups can often be more effective.

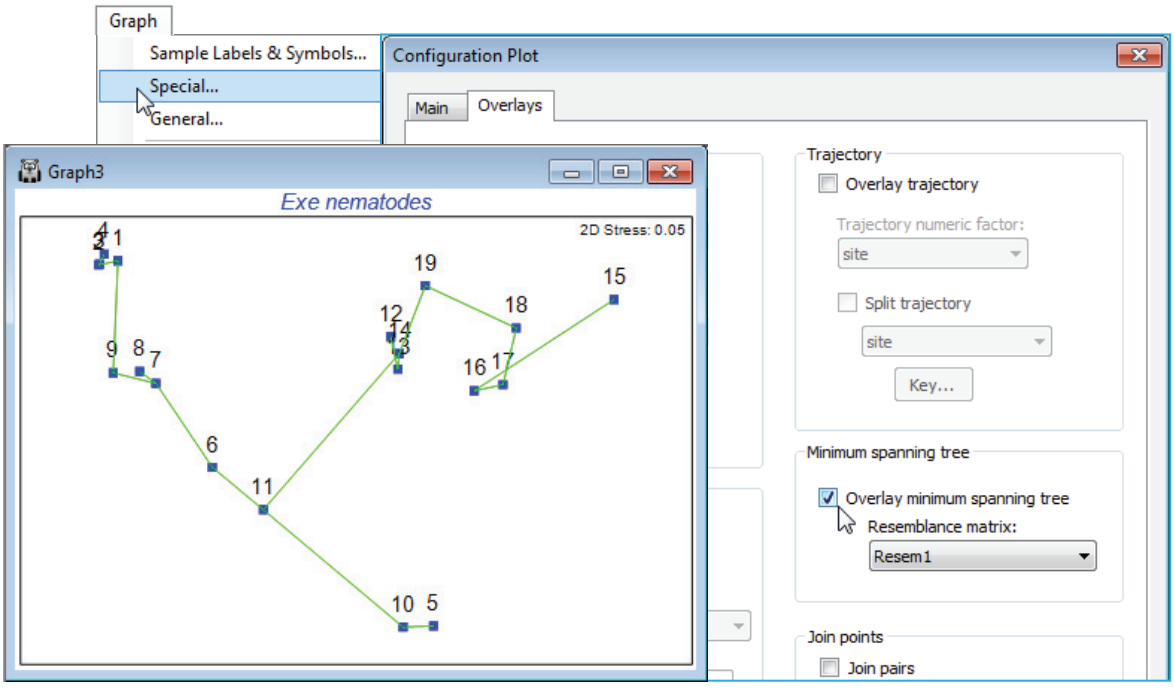
v7



Minimum spanning tree (MST)

v7

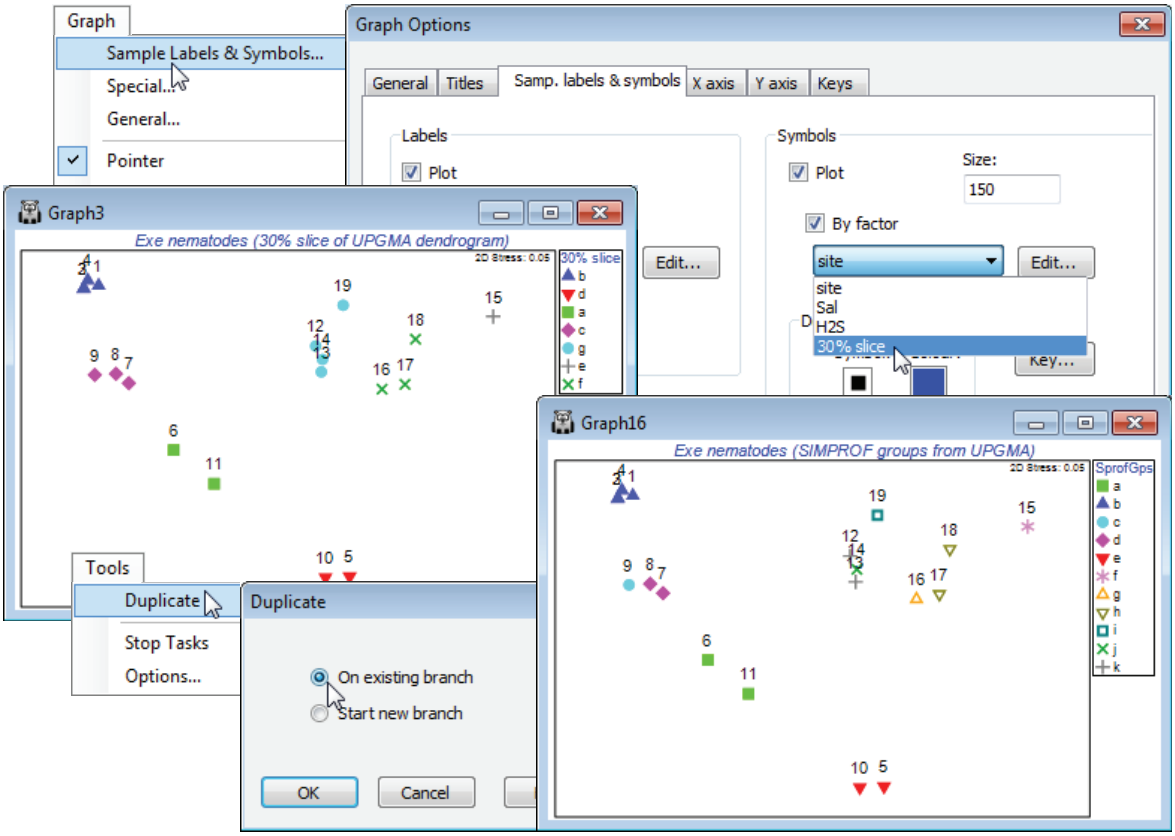
The minimum spanning tree (MST, see CiMC for the Gower and Ross, 1969 reference) is used in the same way, to identify parts of the configuration which do not fully represent the underlying (dis)similarity matrix. All samples are connected on the ordination by a single line, which may branch but never forms a closed loop, chosen to be of minimum 'length' – not using the distances on the ordination but summing the associated dissimilarities from the resemblance matrix. Again, whilst generally the line coincides with a natural joining up of the points in the 2-d ordination, the occasional unnatural links (15 to 16, 19 to 14) show the minor stress in the low-d approximation. You should avoid, of course, taking both MST and Join pairs options at the same time!



Linking MDS plots to cluster analysis

The 2-d ordination plot shows a clear separation of the nematode assemblages at these 19 sites into 5 groups (which CiMC, Chapter 11, shows can be related to sediment properties such as the median particle size, anoxic layer depth and interstitial salinity). Another useful check on the adequacy of the 2-d approximations represented by both MDS and cluster analyses, to the real high-d structure (there are 140 species variables in the Exe nematode abundance data sheet!), to examine the MDS and dendrogram results in combination, and there are at least two ways of doing this.

Firstly, the clusters that are defined for a fixed similarity slice through the dendrogram can be put into a factor, with levels defining the different groups, as seen in Section 6. This factor can then be displayed as differing symbols on the MDS, and the agreement noted. A variation of this which may often be preferable is to use the factor (or factors) created by the series of SIMPROF tests which accompany the particular clustering method (or methods), defining group structures for which there is some statistical support. For the current Exe ws workspace, a similarity slice should already exist as the factor *30% slice* from the dendrogram Graph1. (If not, recreate it, with Graph1 as the active window, by **Graph>Special>**Slicing:(✓Show slice)>(Resemblance: 30) & (**Create factor**>Add factor named: 30% slice). From the MDS (Graph3), take **Graph>Sample Labels & Symbols**>(Symbols✓Plot)>(✓By factor: 30% slice) & (Labels✓Plot)>(✓By factor: site). (When plotting both labels and symbols, note that the symbol is centred on the point, with the label above. On its own, either is correctly centred). You might also like to re-run the **Analyse>Cluster** routine on Resem1, for one of the clustering methods discussed in Section 6, defining SIMPROF groups by a factor which is again used as symbols on the MDS. (Use **Tools>Duplicate>**(•On existing branch) on Graph3 to create a copy of the MDS so you can juxtapose the differing factor selections. You may also want to re-order the key by clicking on it and using ↑ and ↓ repeatedly in the resulting Key dialog). You will see that a finer distinction of sites into clusters is obtained with SIMPROF, implying there is statistical evidence for interpreting such fine-scale groups – but the ability to slice the dendrogram at some arbitrary coarser similarity (or to define fewer groups with the flat-form clustering, *kRCluster*) may still be a justifiable approach for a practical application of site grouping.



A good example of the comparison of different clustering methods and the SIMPROF groups they generate is seen in CiMC, Fig. 3.10, for the zooplankton data in the Bristol Channel ws workspace. Section 6 produced a range of SIMPROF factors: *SprofGps*, *Flexbeta*, *Single* & *Complete* linkage (agglomerative hierarchical), *Unctree* (divisive hierarchical) and *Flat R* (flat-form) clusters, which you may wish to represent as symbols on duplicated MDS plots, in a similar way to the above.

v7

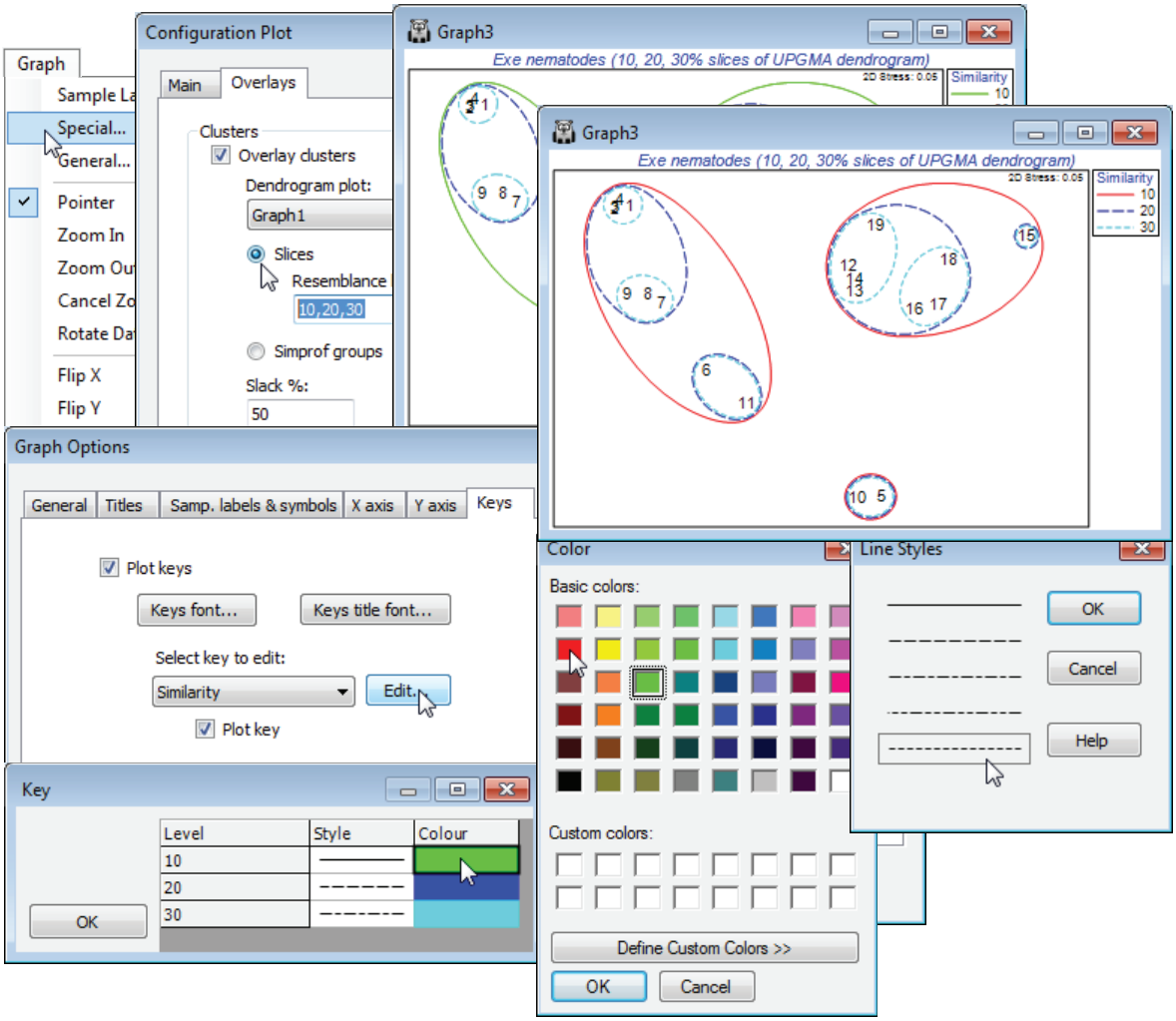
Cluster overlays on MDS plots

v7

The second way in which a cluster analysis can be displayed on a 2-d MDS plot, to aid assessment of the level of agreement, is to draw smoothed envelopes around each of the cluster groups, either for one or more slices at arbitrarily chosen similarity levels (drawn with different line colours and line types) or, as a new option in PRIMER 7, with a previously-defined SIMPROF grouping from a hierarchical dendrogram. The envelopes are smoothed *convex hulls* of the points they enclose and a *slackness* parameter determines the smoothness of the enclosing line (how loosely it is drawn round the points in that group). The default of (Slack %: 100) results in a high degree of smoothing, and thus larger envelopes, with (Slack %: 0) giving the tightest enclosing polygon (the convex hull).

For the Exe nematode MDS (Graph3), remove the symbols that are displayed with the 30% slice by **Graph>Sample Labels & Symbols**, unchecking the (Symbols: ✓Plot box) to leave only the site labels. Take **Graph>Special>Overlays**>(Clusters✓Overlay clusters)>(Dendrogram plot: Graph1) & (•Slices>Resemblance levels: 10, 20, 30) & (Slack %: 50). Experiment with the slack parameter and change the colours and line types with the Keys tab in the Graph Options dialog, i.e. **Sample Labels & Symbols>Keys>**(Select key to edit: Similarity)>**Edit** and double-click on a colour or line style box to get the same options as for factor Keys, seen previously. You might also like to add envelopes for the SIMPROF groups created by the cluster run you chose on the previous page – by taking (Clusters•Simprof groups) rather than (Clusters•Slices) on the Overlays tab.

v7

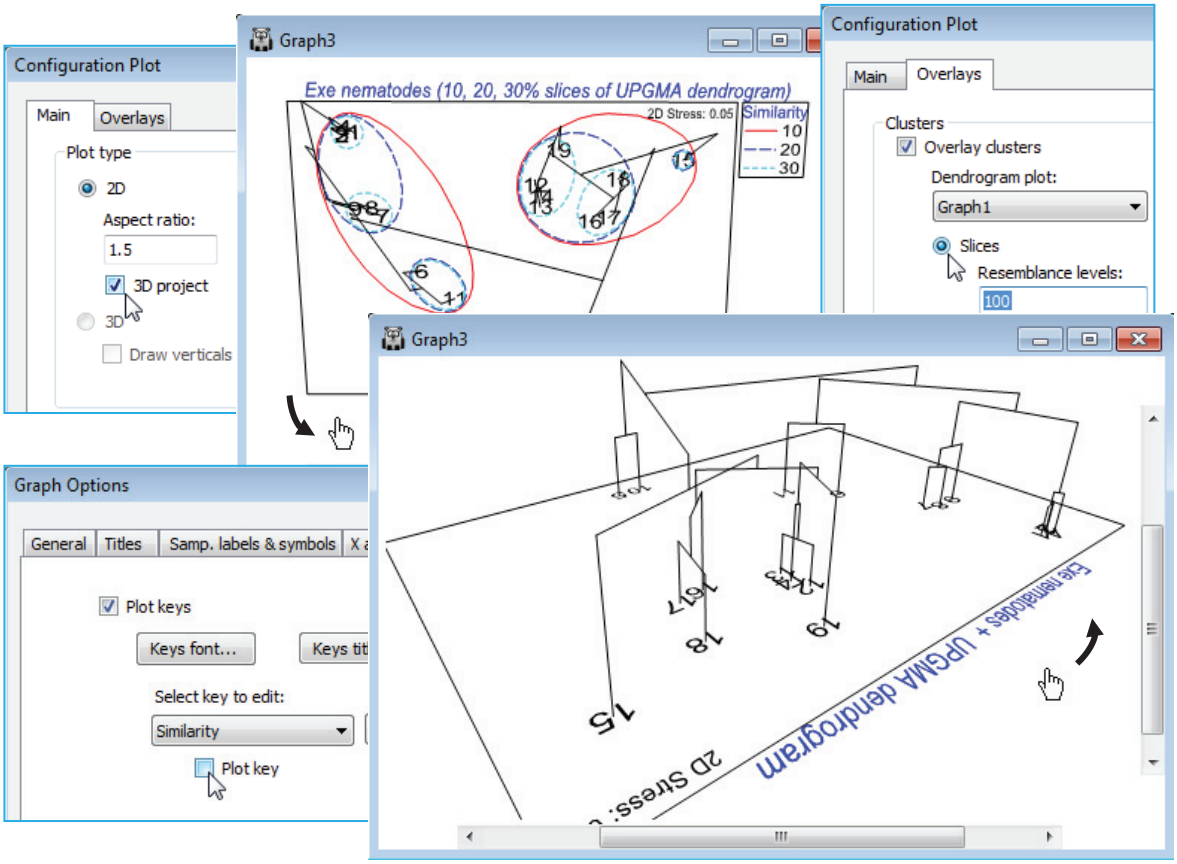


Attractive though such envelopes are, in guiding the eye to cluster groupings, there can be a lack of clarity as to which points fall in which envelopes when boundaries overlap, sometimes exacerbated by too high a slack parameter (chosen in the interests of producing very smoothed curves). This is also the reason why the envelope overlay operation is not offered for a general *a priori* factor, defining a one-way layout of samples, for which there is quite likely to be substantial overlap of some groups. At least with SIMPROF groups or dendrogram slices, if the MDS and cluster analysis are substantially in agreement, the likelihood is that most cluster groups will occupy a discrete region of the MDS space and lightly smoothed convex hulls will tend not to overlap too often. And *a priori* factors can always be simply and unambiguously displayed using different symbols.

Dendrogram & 2-d MDS in a 3-d plot

v7

Rather than creating a small number of (arbitrary) slices through a dendrogram, superimposed on a 2-d MDS, a further new feature in PRIMER 7 is the ability to draw the dendrogram as the third dimension in a 3-d plot of the 2-d MDS (or any ordination). As with all 3-d plots (see shortly), the graph is then rotatable, to allow good visualisation of the 2-d MDS in juxtaposition with the full dendrogram. This is again accessed from the **Graph>Special** dialog. With the current 2-d MDS for the Exe nematode data (Graph3), under the **Main** tab take (Plot type•2D)>(✓3D project). Under the **Overlays** tab, it is also necessary that you select (Clusters>✓Overlay clusters) and take either (•Slices) or (•Simprof groups), with the correct plot dendrogram specified, e.g. (Dendrogram plot: Graph1) for the former. From the previous page, the option (Resemblance levels: 10, 20, 30) will already be implemented, giving the first plot below, but if it is considered unnecessary to duplicate the envelopes on the 2-d plot as well as adding the full dendrogram, the envelopes can be switched off, effectively, by taking (Resemblance levels: 100). To get precisely the second plot below, there are some further minor steps: the title has been changed from the **Titles** tab on the standard Graph Options dialog box; the unneeded Similarity key is removed by unchecking the (✓Plot key) box on the **Keys** tab of Graph Options; the plot has been zoomed with **Graph>** (or right-click) **Zoom In** or the  icon (this is often a useful step with a 3-d plot – note the scroll bars which allow the figure to be appropriately centred). Finally, the plot is rotated with **Graph>Rotate Axes** or the  icon, then clicking, holding and dragging with the cursor, which is now a hand.



3-d ordination plots & axes selection

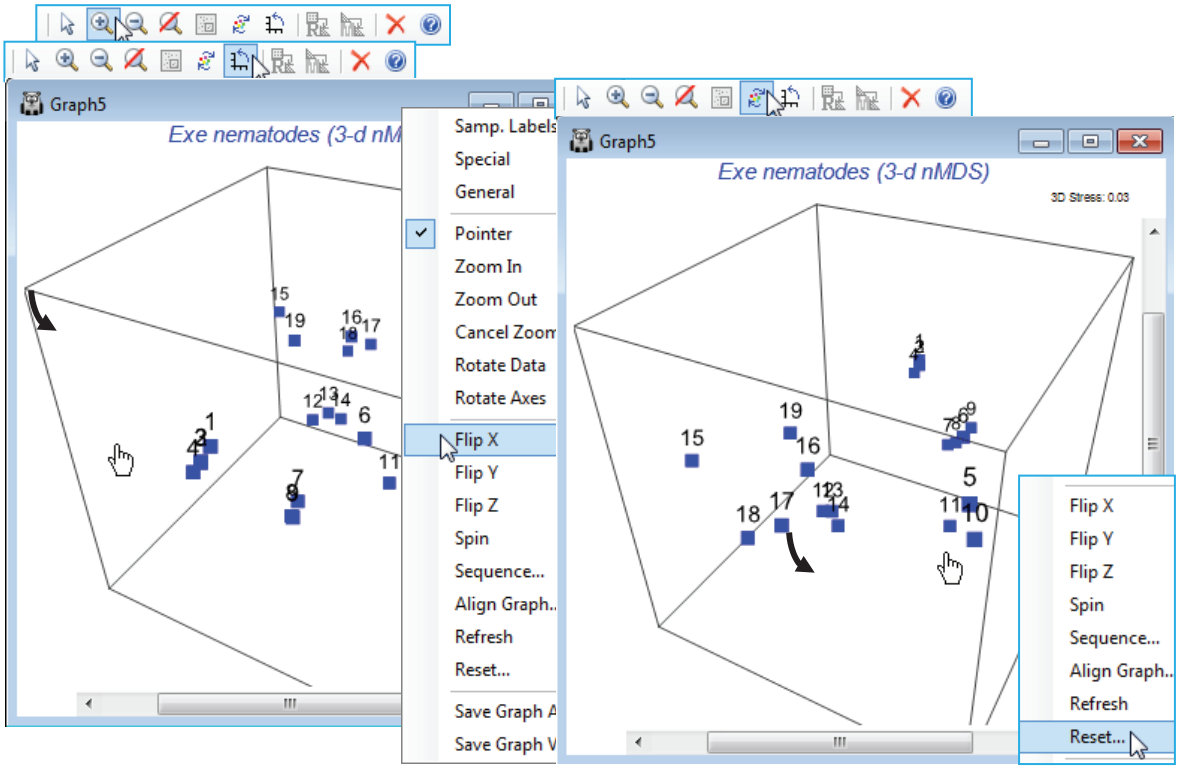
v7

As we have seen for the *n*MDS run on the Exe nematode data, in the first graphic of this Section, a default MDS run produces both 2-d and 3-d configuration plots, and their associated Shepard diagrams, placed into a multi-plot (Multiplot1). A different plot window is created for the 2-d than the 3-d solution because the operations are entirely separate. This is in contrast with PCA (Section 12) or PCO (in the PERMANOVA+ add-on) where the 2-d solution is just the first two axes of the 3-d solution, and a single plot window is sufficient to capture all the possibilities. There, the **Graph >Special>Main>**Axes section will allow a choice of which two or three of the (many) PCA axes are selected for plotting, in either the specified (Plot type•2D) or (Plot type•3D) configuration. For MDS plots, which can in PRIMER 7 also be produced for many dimensions, each new dimension requires a different iterative process and, for example, the 2-d solution will not be the first two axes of the 3-d solution. Hence the plot windows for different dimensions are listed separately, though the option to select different combinations of axes from each, for a 2- or 3-d plot, remains the same.

Rotate axes or rotate/flip data

The 3-d Exe nematode *n*MDS plot should be in the current workspace as Graph5 under Multiplot1 (if necessary create it again, with default options, by **Analyse>MDS>Non-metric MDS (nMDS)** on the Bray-Curtis resemblance matrix for a square-root transform of the data in Exe nematode abundance). Uncheck the (Symbols✓By factor) box on the (right-click) **Samp. Labels & Symbols** dialog and again **Zoom In** and take **Rotate Axes**, turning the 3-d box to properly see the position of these points in 3-d space. Note the distinction here between **Rotate Axes** (the icon) and the **Rotate Data** option (the icon). For a 2-d plot there is no benefit in rotating the axis box to produce a slanting rectangle surrounding the points(!), so this is not implemented. Rotating the points within the rectangular box can occasionally be useful, in order to line them up with a similar ordination for example, and bearing in mind the arbitrary orientation of points – this uses **Rotate Data**. In 3-d, rotating the box is beneficial because it allows a static plot of the points to be viewed dynamically from a range of angles so that the 3-d structure can be properly appreciated – this uses **Rotate Axes** and is available for any 3-d plot. Specifically for 3-d MDS plots, where the relation between the axes of the box and the orientation of points is arbitrary, **Rotate Data** (within a static axis box) is also available, along with reflecting the points in the axes with **Flip X**, **Flip Y**, **Flip Z**. Generally, rotating data will not be as useful as rotating the axes themselves but it can be important where, for example, an MDS plot is being referred to a physical layout of sampling sites in a 3-d medium. Try out the various combinations of rotation/flipping for Graph5, bearing in mind that you can always restore the original relationship of the points to the 3-d box by (right-click) **Reset**.

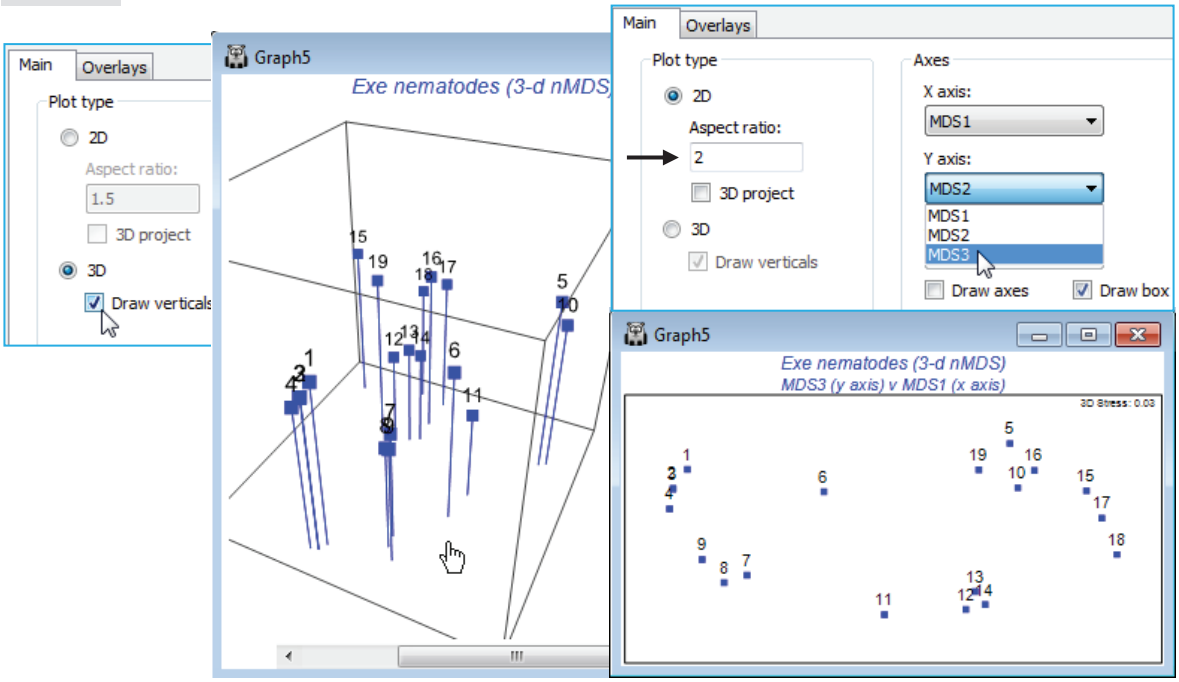
v7



Drawing verticals for 3-d plots

v7

Whilst it is relatively easy to visualise a 3-d plot by rotation of the axes – and PRIMER 7 is now able to output digital video of such rotations (as *.mp4, see below) – it is much harder to produce a convincing 3-d plot for static reproduction in 2-d, e.g. for a printed publication. PRIMER 7 tries to aid this in two ways: by giving any text on the plot a pseudo-perspective, e.g. with site or axes labels shrinking with distance into the plot, and providing an option on the **Graph>Special>Main** tab to (Plot type•3D)>(✓Draw verticals), which drops vertical lines onto the base plane of the box. For a limited number of samples this might help to fix the relative depths into the plot of the points. Alternatively, 3-d MDS axes could be viewed in 2-d, a pair of axes at a time. This uses, e.g. for the 3-d MDS plot of the Exe data (Graph5), **Graph>Special>Main>**(Plot type•2D) & (Axes>X axis: MDS1 & Y axis: MDS2) then (X axis: MDS1 & Y axis: MDS3) and possibly (X axis: MDS2 & Y axis: MDS3). It might be sensible to duplicate the first of these plots, with **Tools>Duplicate>**(•On existing branch) so that two (or all three) of these plots can be viewed together in the workspace. The same idea could be used to generate two or more of the four possible 3-d plots obtainable from the axes of a 4-d MDS (or PCA/PCO), e.g. (X axis: MDS2 & Y axis: MDS3 & Z axis: MDS4), etc.



v7 The plot shown above right (of MDS3 against MDS1) also uses the capacity, noted earlier, to alter the box shape for 2-d plots, by setting (Aspect ratio: 2) rather than the default of 1.5. In fact, rather little is to be gained by a 3-d MDS solution in this case since the 2-d stress of 0.05 is already low, the 3-d solution takes it down only slightly (to 0.03), and the 2-d approximation which the MDS solution represents is unlikely to mislead at all. Save and close the Exe ws workspace.

(W Australia fish diets)

Re-open the workspace WA fish ws from C:\Examples v7\WA fish diets, which will provide an example where a 3-d MDS is necessary to get a reliable feel for the multivariate structure. The workspace WA fish ws should contain a data sheet Data2 of the dietary assemblage of (pooled) guts of 7 marine fish species from W Australia, with uneven numbers of replicate pools per species. If not available, open WA fish diets %vol.pri; the Data2 sheet simply excluded three samples, A9, B3, B4, which had much lower total gut content than the remaining 65 pools – see Section 3 on deselecting samples. Take **Pre-treatment>Standardise>**(Standardise•Samples) & (By•Total) and transform the result with Pre-treatment>Transform(overall)>(Transformation: Square root) and **Analyse>Resemblance>**(Measure•Bray-Curtis similarity), renaming this WA B-C sim.

Higher-d & scree plots (WA fish diet)

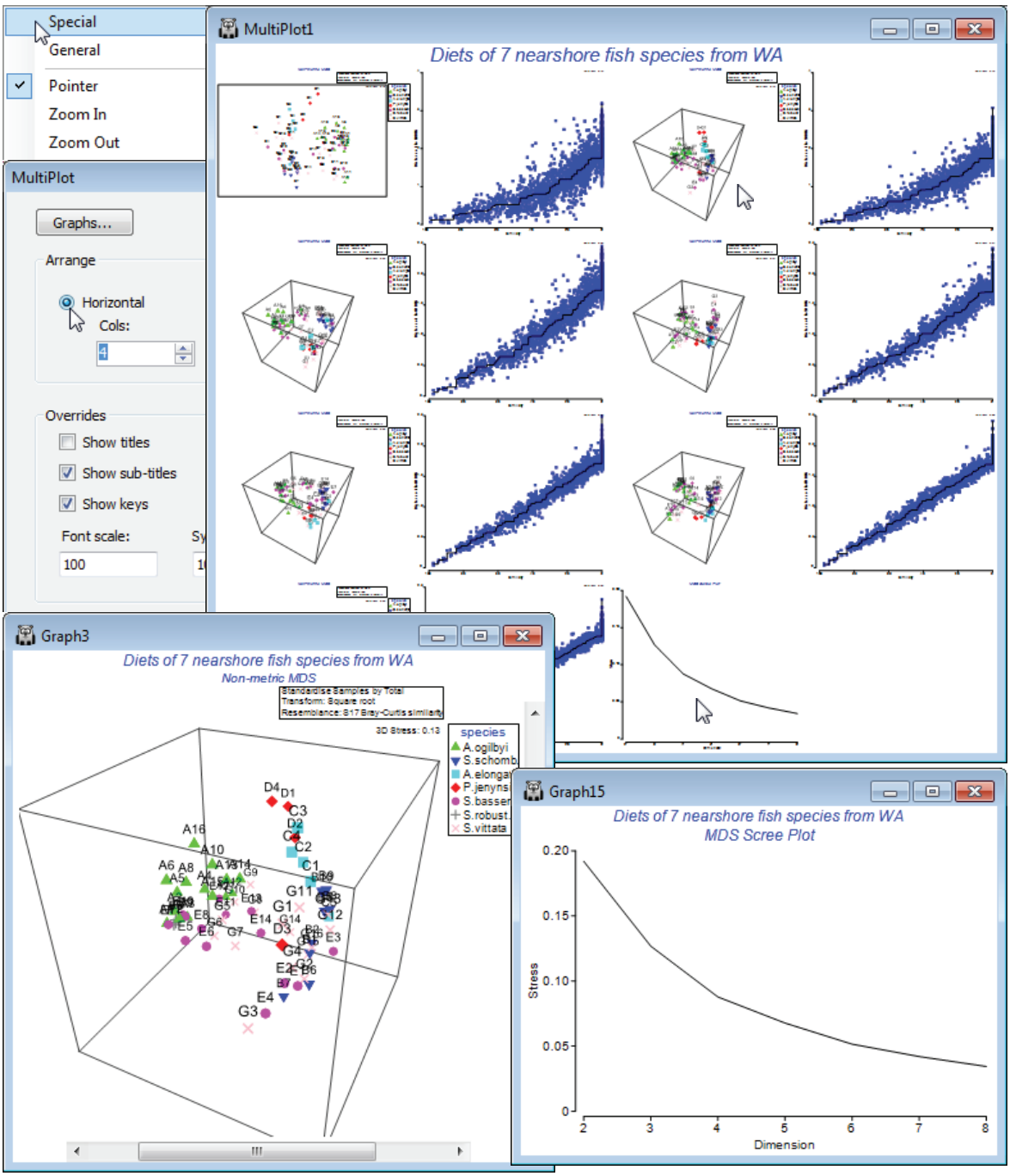
Run *n*MDS at a wider range of dimensionalities than the default and ask for a *scree plot*: **Analyse> MDS>Non-metric MDS (mMDS)>**(Min. dimension: 2) & (Max. dimension: 8) & (✓Scree plot), taking the other defaults (i.e. including Shepard plots). On the resulting multi-plot, Multiplot1, take **Graph>Special**>(Arrange•Horizontal)>(Cols:4) to make sure the MDS plots and their associated Shepard diagrams are arranged in two pairs across each of the 4 rows, finishing with the scree plot.

v7

Clicking on the 2-d configuration you can see that though there are some clear differences in the dietary assemblages for some fish species, the stress of the plot is high (0.19), reflected in the large scatter in its Shepard diagram. The 3-d (& 4-d) solutions are noticeably better, with the stress falling to 0.13 (and then 0.09); the scree plot shows this initially quite steep decline (stress must always decline as the number of dimensions available, to display the relationships among samples, increases). Of course, solutions in higher than 3-d can only display three co-ordinates at a time, so the configurations in the multiplot all show only axes MDS1, MDS2, MDS3. Whilst sets including some higher axes could be viewed for the ≥4-d solutions, e.g. MDS1, MDS2, MDS4, the primary visualisation here should be the 3-d MDS, which has a stress at 'acceptable' levels (see discussion on stress levels in CiMC, Chapter 5). Click on the 3-d plot in the multiplot (Graph3) and **Zoom In**, remove the uninformative labels by unchecking (Labels✓Plot) from the **Samp. Labels & Symbols** tab, remove the history by unchecking (✓Plot history) on the **General** tab, edit the title and delete the subtitle by removing its content on the **Titles** tab. You may also want to change the colour or symbol for some of the species by clicking on the key, which sends you to the Key dialog box, as previously seen. Finally rotate the 3-d axes manually by **Graph>Rotate Axes** (or the icon), to note how some fish species (e.g. *S. vittata*) feed widely across the dietary space and others are more specific (e.g. *S. robustus*). Testing of these dietary differences requires ANOSIM (Section 9).

v7

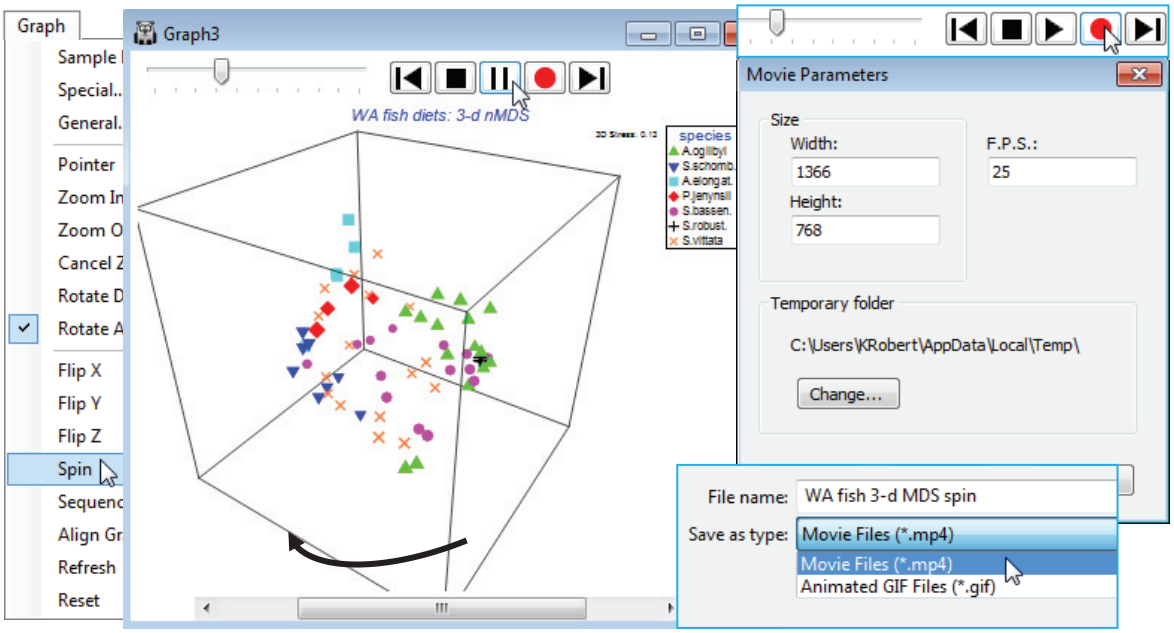
v7



Spinning a 3-d MDS & capture in a movie file

v7

The 3-d MDS plot box can be made to rotate horizontally, automatically (from whatever vertical perspective view it is manually set to), by **Graph>Spin**. This adds a top line of controls for the animation: to the left is a slider control which sets the speed of rotation (click and drag to the right for higher speed), followed by a bank of standard video-type controls, the key ones of which are to start the automatic spin with ▶ , which then changes to a pause button ⏸ and an end to the **Spin** routine (and removal of the controls) is achieved by the ■ button. PRIMER 7 has also introduced the ability to capture three automatic animations of this type (the other two, seen later, are a sequence animation and evolution of an MDS construction) to a digital 'movie' file with an *.mp4 or animated *.gif format, so that 3-d plots and the other animations can be embedded in presentations (or perhaps supplementary material for publications) in the same way as static graphs. Recording is launched from the ● button, which gives a Movie Parameters dialog box specifying: pixel sizes for the images, the default being (Width: 1366)&(Height: 768); the number of frames per second, default (F.P.S.: 25); and an option to change the temporary folder used in the recording. The usual Save As dialog box then allows the directory and filename for the movie file, and the file type (*.mp4, animated *.gif) to be specified, and recording starts with ▶ and finishes with ■. Long recordings should be avoided: the movie files will get very large, very quickly!



(Re)save the workspace as WA fish ws for use in later in this section, and close it.

(Morlaix macrofauna, Amoco-Cadiz oil spill)

Some data sets have a natural sequence to their samples, usually a time series (though this can be a single spatial gradient), and it is usually a good idea to join the points on an ordination plot in that order – then referred to as *trajectory plots*. This is illustrated by a 'classic' data set of soft-sediment macrofaunal assemblages in the Bay of Morlaix at a single station (*Pierre Noire*), sampled at about 3-monthly intervals over the period April 1977 to February 1982, covering the event of the Amoco-Cadiz oil tanker wreck in March 1978. The spill occurred some 40 km from the Bay itself but oil slicks reached this area and there is a clear signal of community change in the sampling times after the spill, with a partial recovery over the next 3 years – see the MDS below. There are 21 sampling times (A-U) over the 5 years, with the oil impact occurring between samples E and F. These data are from Dauvin J-C 1984, *Ph.D. thesis, Univ Pierre et Marie Curie, Paris*, though a dynamic view of the continuation of this time series (Dauvin J-C 1998, *Mar Pollut Bull* 36; Thiébaut *et al* 2012, *Conference: Time series analysis in marine science*, Logonna Daoulas, France) is also seen later.

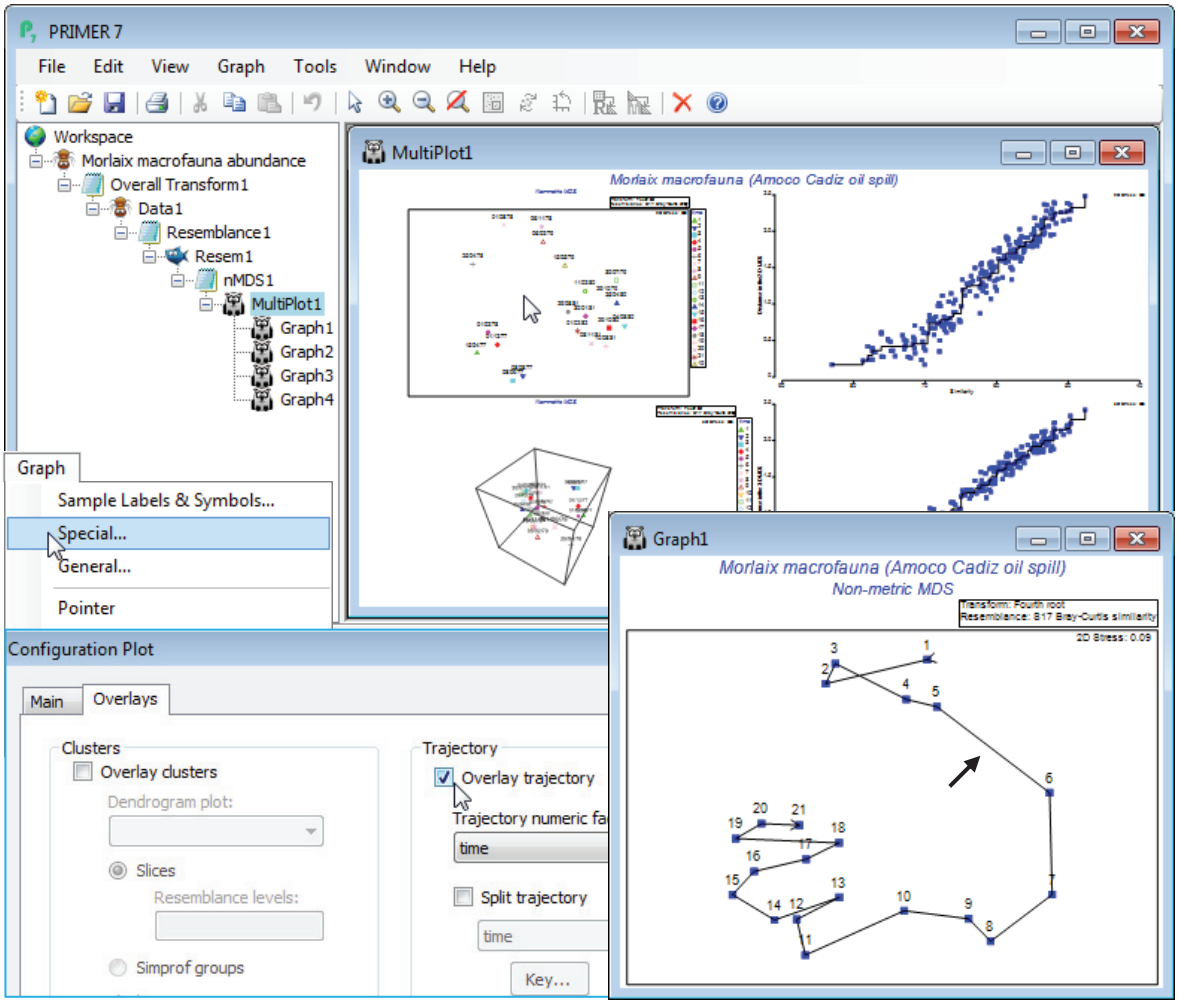
Overlay trajectories

A trajectory joining points on an ordination is simply added using **Graph>Special>Overlays** and taking the (✓Overlay trajectory) option and supplying a factor name for the trajectory sequence, which of course needs to have purely numeric entries. These do not need to be the integers 1, 2, 3, etc – the rank of the numbers will be used to dictate the connection order. Any blank entries in the factor will be ignored – these points will not be part of the trajectory, but neither do they break a connection to give multiple trajectories. In order to produce the latter, a new (and very useful!) feature in PRIMER 7 is to tick a (✓Split trajectory) check box and supply a second factor name. This is now a categorical factor with differing entries determining the separate trajectories to be drawn, in the same sequence order for all such groups, specified by the first (numeric) factor. A typical example would be an MDS plot of a time series of samples taken at a number of sites, the first factor determining the sampling year (or month or day etc) and the second factor the site name, allowing the time progression to be tracked more clearly on the ordination, in parallel for the sites. A **Key** button gives access to a standard dialog box which sets the symbol, line type and colour (the same for both symbols and trajectory) for the different groups.

v7

To demonstrate a single time trajectory, open data file Morlaix macrofauna abundance in directory C:\Examples v7\Morlaix macrofauna, and carry out an *n*MDS in standard fashion, much as for the previous example, except with no sample standardisation and use of a fourth-root transform prior to the Bray-Curtis calculation. Use all the default settings for *n*MDS to produce the 2-d plot (Graph1), which has low stress of 0.09 (bearing in mind that the original data has 251 species dimensions, many of which enter the similarity computation because of the severe fourth-root transformation!). As above, harmonise the symbols, i.e. on the **Samp. labels & symbols** tab remove (Symbols✓By factor) and add (Labels✓By factor: time) – also for the below, the **General**>(Overall font scale) was increased, and you may need to **Rotate Data** or **Flip** an axis to obtain the configuration shown.

Graph>Special>Overlays>(Trajectory✓Overlay trajectory)>(Trajectory numeric factor:time) then adds the time trajectory and greatly clarifies the interpretation. The first 5 points represent a year's seasonal cycle prior to the spill (the impact time indicated by the arrow), after which community structure changes strongly over the next year (4 points) before an apparent partial recovery towards the initial community, with the scale of the three seasonal cycles evident for the last three years.



Sequence animation, captured in 2- & 3-d

v7

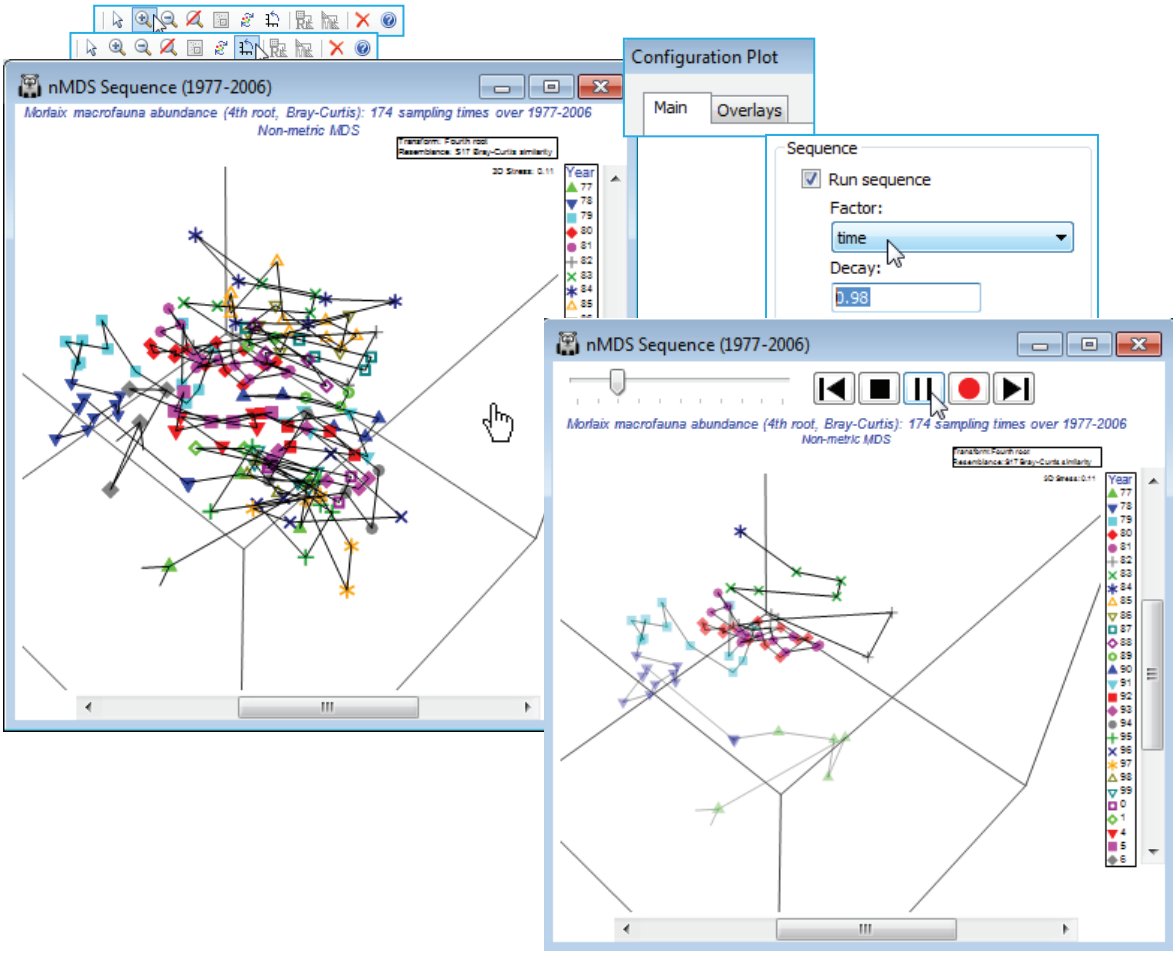
The time pattern above is evident from simple addition of the trajectory but there are occasions when a confused MDS plot results from a longer sequence of points, especially if the multivariate structure tracks back on itself, repeating earlier states. Viewing the 3-d MDS is often helpful, and a trajectory can be added there, in exactly the same way. However, it can also help to elucidate the sequential pattern dynamically. PRIMER 7 has thus added *sequence animation* as a further option for 2- or 3-d ordination plots. This adds the samples to the ordination plot in sequential (e.g. time-ordered) fashion, including the joining lines of the trajectory if that is selected (as it usually would be). The earlier points and lines in the sequence gradually fade out as the later ones are plotted, allowing a more uncluttered view of progression of the multivariate structure. This is accessed by **Graph>Special>Main>**(Sequence✓Run sequence), supplying the numeric factor which gives the temporal (or other sequence) order and specifying a Decay parameter (in the range 0 to 1), which controls how quickly the earlier parts of the sequence fade. The default of 0.9 corresponds to a rather slow fade so that most of the earlier points from a short sequence remain visible throughout. Other contexts may need an even slower (0.98) or faster (0.5) decay, both these being used below.

The sequence animation can be run with or without capturing the dynamic display in digital movie form. From the **Main** tab, on entering the ordering factor and decay, and pressing **OK**, the standard video controls are displayed, and these operate just as described two pages ago (for **Graph>Spin** on a 3-d MDS plot): ▶ to start or pause, ■ to stop, with the slider varying the animation speed. A further option available here is to take one sequence step forwards or backwards with ▶| or |◀. If ● has initially been pressed, recording starts with ▶ and REC appears; stop by ■ . If correct entries have already been set in a previous run, repeats need only a selection of **Graph>Sequence**.

To demonstrate this sequence animation, **File>Open** the plot file nMDS Sequence (1977-2006).ppl into the current Morlaix workspace. This is a PRIMER binary-format file of a 3-d MDS plot for a much longer run of years of grab sampling for benthic macrofauna at station *Pierre Noire* in the Bay of Morlaix, see Dauvin and Thiébaut references above. The data matrix is not accessible from this file – it is used here only as a dynamic demonstration of the (static) MDS plot in the Thiébaut *et al* analysis – but it is worth noting that a PRIMER *.ppl file does carry around with it enough of the background information on the samples (the factor sheet) to allow the ordination configuration to be annotated in different ways, rotated etc. So this could be a flexible format for distributing plot results to others (who have access to PRIMER 7) but where the data itself cannot be released.

Take **Zoom In** (and centre with the scroll bars) and manually **Rotate Axes**, but the pattern is still a little confused. Does the benthic community approach its 1977 pre-oil impact state (if that one year is representative of this!), or are more regional or global-scale processes of environmental change at work? A slightly clearer demonstration of the time course is produced by **Graph>Special>Main** >(Sequence✓Run sequence)>(Factor: time)&(Decay: 0.98) and ▶. You will find that rotating the plot while the sequence unfolds is of some help, and perhaps varying the speed with the slider. The plot captures not just the inter-annual trends but also (if carefully watched) the scale of the seasonal cycle in relation to those trends (there is an average of nearly 6 sampling times across each year).

v7

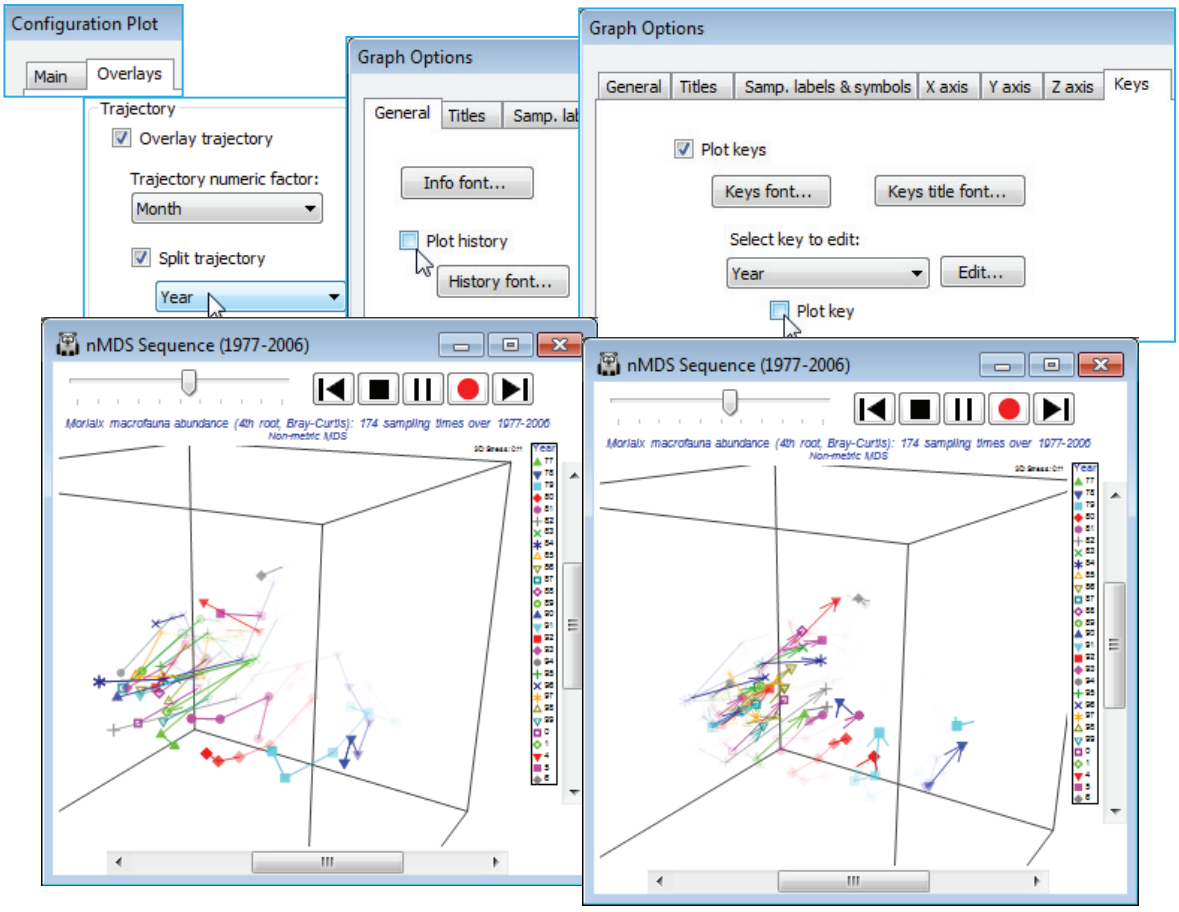


Trajectories split & then sequence animated

In fact, if the animation is repeated with trajectories running across months, separately drawn for each year, using the option discussed earlier to split the trajectories, then the scale of the seasonal cycle can be made visually apparent in a dynamic way. Firstly, take **Graph>Special>Overlays>** (Trajectory✓Overlay trajectory)>(Trajectory numeric factor: Month)&(✓Split trajectory>Year), then you might like to remove extraneous information from the plot, such as the History box on the **General** tab and the line keys on the **Keys** tab, unchecking (✓Plot key) under the second entry for (Select key to edit: Year). Then the animation **Graph>Special>Main>**(Sequence✓Run sequence)> (Factor: Month)&(Decay: 0.5) gives a dynamic display of the seasonal cycle running in parallel across the years.

v7

Save the workspace as Morlaix ws and close.

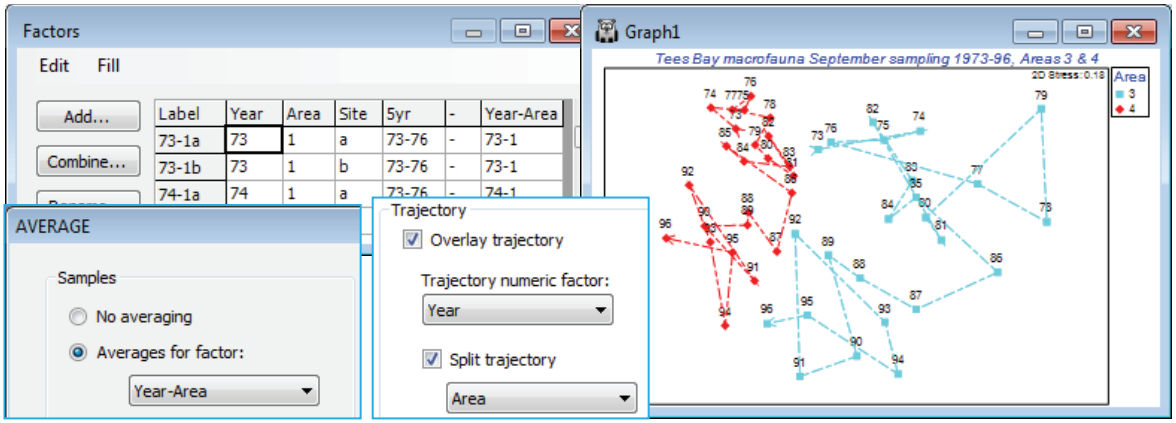


(Tees Bay macrofauna time series)

A clearer, static, example of the benefits of split trajectories on ordination plots is provided by macrofauna data from Tees Bay, collected and analysed by the Brixham Laboratory, SW England, in (amongst other locations) four sub-tidal areas, 1 to 4, spanning approximately 10km of the N Sea coast of the UK, with two sites (a/b, c/d, …) in each area, and these data result from pooled grab samples at each site. The samples in this case were collected in September, annually from 1973 to 1996 (Warwick *et al*, 2002, *Mar Ecol Prog Ser* 234: 1–13).

Open Tees macrobenthic abundance.pri in C:\Examples v7\Tees macrobenthos, and the factors by **Edit>Factors**, and create a combined *Year-Area* factor. (Do this by **Add>**(Add factor named: -), put a hyphen in the first cell, highlight the column by clicking on the factor label -, and **Fill>Value**, then take **Combine** and move Year, - and Area names to the Include box; this construction of a combined factor was illustrated in Section 2). Now fourth-root transform the data, average over the sites with **Tools>Average>**(Averages for factor: Year-Area) and compute Bray-Curtis similarities. Select two areas – one within the immediate environment of the Tees estuary (area 2 or 3) and one outside (1 or 4), see Fig.6.17 of CiMC. For example, **Select>Samples>**(•Factor levels)>(Factor name: Area)>**Levels**>(Include: 3, 4). Now run *n*MDS and on **Samp. Labels & Symbols** set labels as Year and symbols as Area. Finally, on **Special>Overlays** add trajectories of Year, split by Area. Save the workspace as Tees ws – it is needed for hypothesis tests in Section 9 – and close it.

v7



Matching variable sets

v7

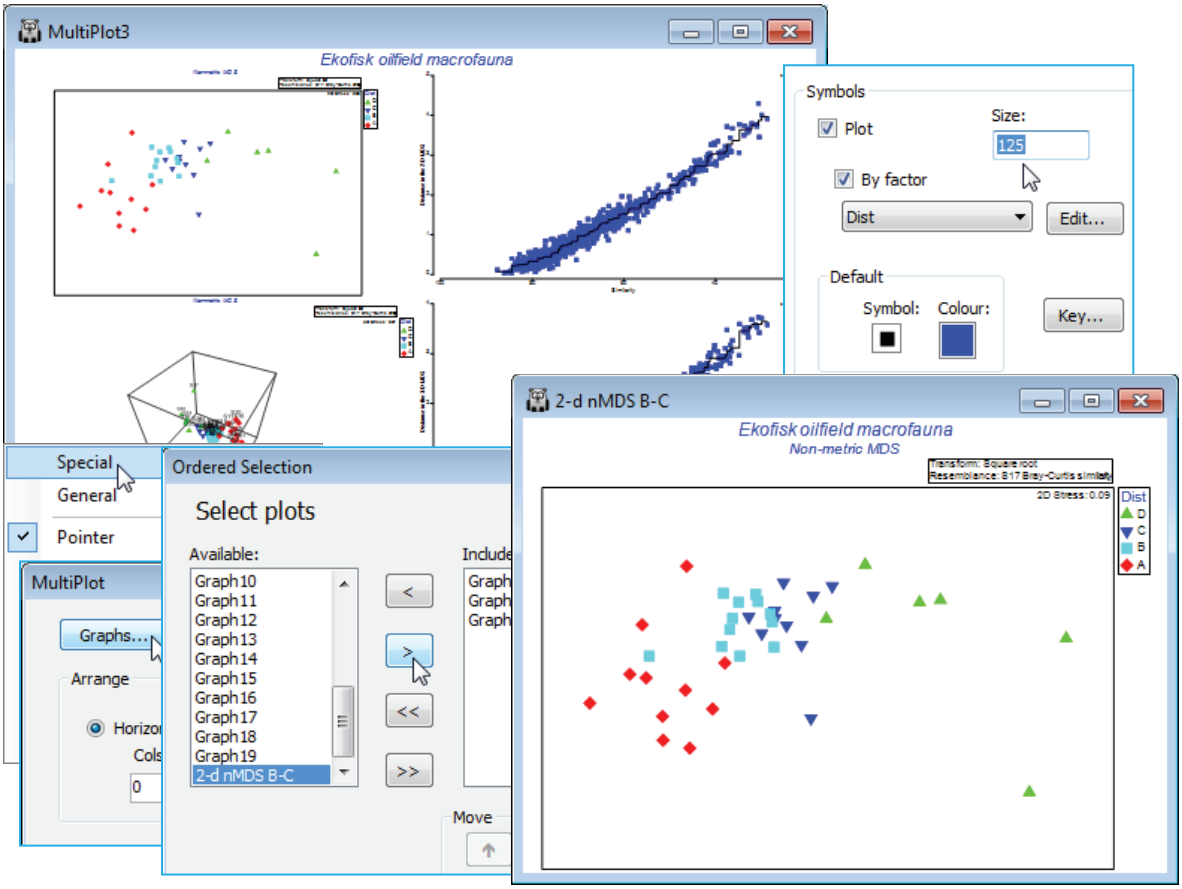
The final two sections of the **Graph>Special** dialog concern overlaying vectors and 'bubbles' of numeric values, rather than factors, either of the variables (e.g. species) from which the ordination has been created, or taken from an independent worksheet (e.g. of environmental variables) which contains the same sample labels as are found in the ordination. Note the matching principle: for all routines, the selected labels in the active sheet determine which samples are to be analysed, and any secondary data sheet supplied to that routine must include at least that sample label set, with names spelt in exactly the same way. The secondary data sheet can also contain other samples (which are ignored) and, of course, the required samples could be in a totally different order in the two sheets.

(Ekofisk oil-field study)

v7

The superimposition of bubble plots and vectors on an ordination are illustrated for the benthic macrofauna assemblages sampled for 39 sites at different distances from the Ekofisk oilfield, in the Norwegian sector of the N Sea, sampled at one point in time, some years after operations began. This dataset was introduced at the beginning of the manual (Section 1) and the workspace Ekofisk ws last saved in Section 4. If this workspace is not available, open Ekofisk macrofauna counts and Ekofisk environmental from C:\Examples v7\Ekofisk macrofauna, square root transform the former (rename Square root) and compute Bray-Curtis similarities (B-C on sq rt). Run *n*MDS on this to obtain the 2-d ordination, 2-d nMDS B-C. (Note that renaming a graph removes it from the multi-plot – to return it, run **Graph>Special>Graphs** on the multi-plot and take its new name across to the Include list, moving it up to the head of the list in this case. Also, you may need to **Rotate Data** or **Flip X** or **Y** from the right-click menu to obtain the 2-d *n*MDS plot shown below). Already seen in the 2-d *n*MDS display are the four different symbols in the factor Dist, which puts the 39 sites into *a priori* defined groups depending on their distance from the centre of oil drilling activity – D: < 250m; C: 250m to 1km; B: 1km to 3.5km; A: > 3.5km. Use the **Samp. labels & symbols** tab to unclutter the plot by removing the labels and perhaps increase the symbol size, e.g. Size: 125.

The 2-d stress is quite low (0.09) and the ordination thus reliable (with the 3-d plot little improved, as seen in the Shepard diagram). There is a clear pattern of steady change in the community as the rig is approached (left to right). Note that sites within a few hundred metres of the rig (D), and thus close to each other, have quite variable assemblages, but distant sites (A), which can be 16km or more apart, are more tightly clustered – the opposite of expectation if the communities are not impacted by drilling mud disposal. The clear distinction between B and A (confirmed by ANOSIM, Section 9) is good evidence for an impact extending to more than 3 km from the oilfield centre.



Bubble plots of single variables

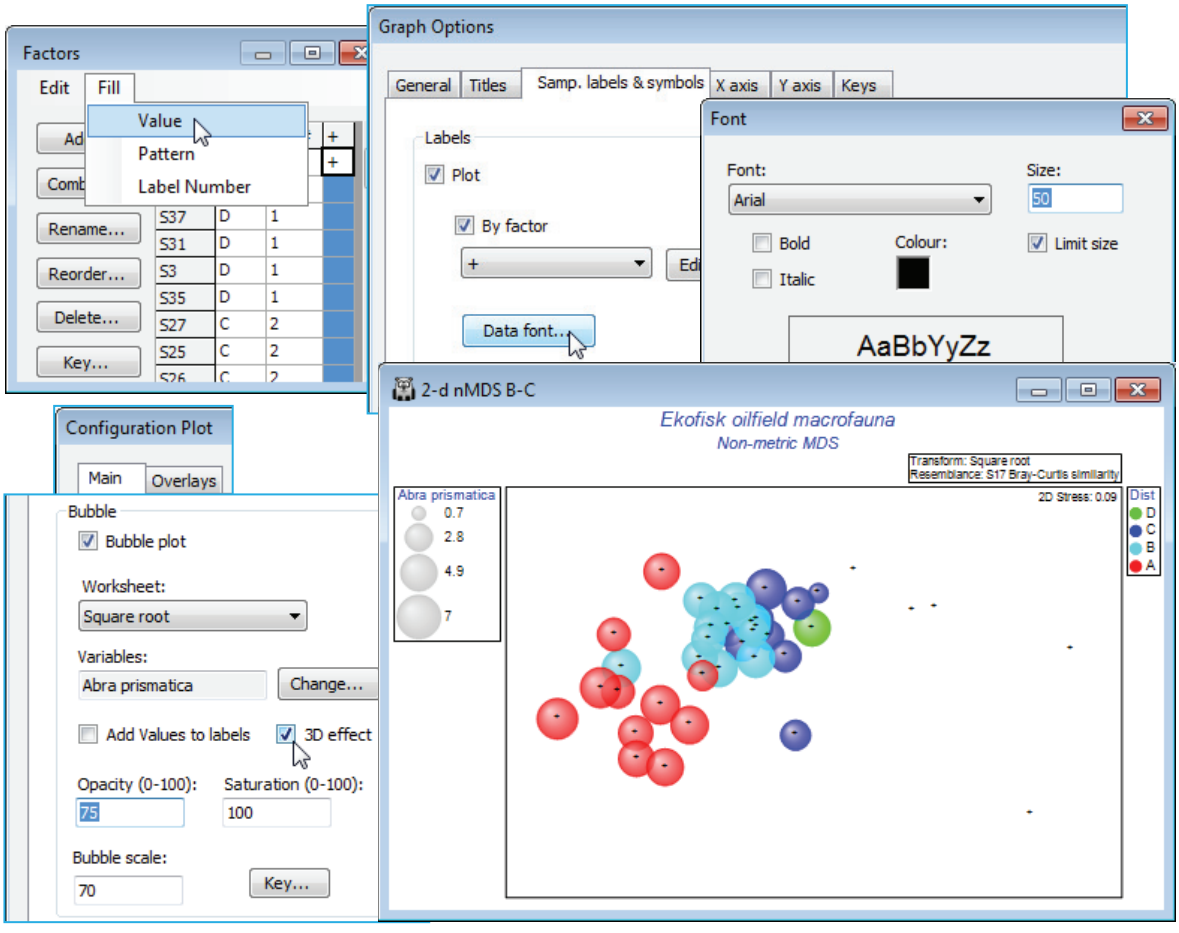
The above MDS draws on all 174 species (with more abundant species given greater weight) but only some species will be responsible for creating the observed gradient – others will be largely 'noise'. The behaviour of a single species over the sites can best be seen by a bubble plot (**Graph> Special**), in which circles are drawn at each point, of size related to the counts at that site. The secondary data sheet here is therefore the (transformed) data matrix itself, Square root.

On the MDS, 2-d nMDS B-C, take **Graph>Special>Main>**(Bubble✓Bubble plot)>(Worksheet: Square root) & (Variables: Abra prismatica), which is the default – alphabetically the first species, but well worth plotting! Taking **OK** now produces a bubble plot for this species, with a scale key (in square root abundance units) which goes down to a vanishingly small bubble for a zero count. Sometimes, for species which are not ubiquitous, it can be helpful to be reminded of where these zero counts are on the plot, by reinstating a label at all points. A useful trick here is to define a new factor which contains just the '+' sign for all samples, plotted at a smaller than default size – with Data font>Size: 50 perhaps, under Labels on the **Samp. labels & symbols** tab. Alternatively, on this same dialog, leave (Labels✓Plot) turned off, but instead use the **Special>Main>**Bubble dialog to (✓Add Values to labels). This is not recommended in this case because there are a number of samples falling close to each other, and especially since square root values with several decimal places will then be added to the plot (this option is best reserved for bubble plots on original scales where it is important to pick out the precise variable value at some points of the ordination).

Bubble colours

v7

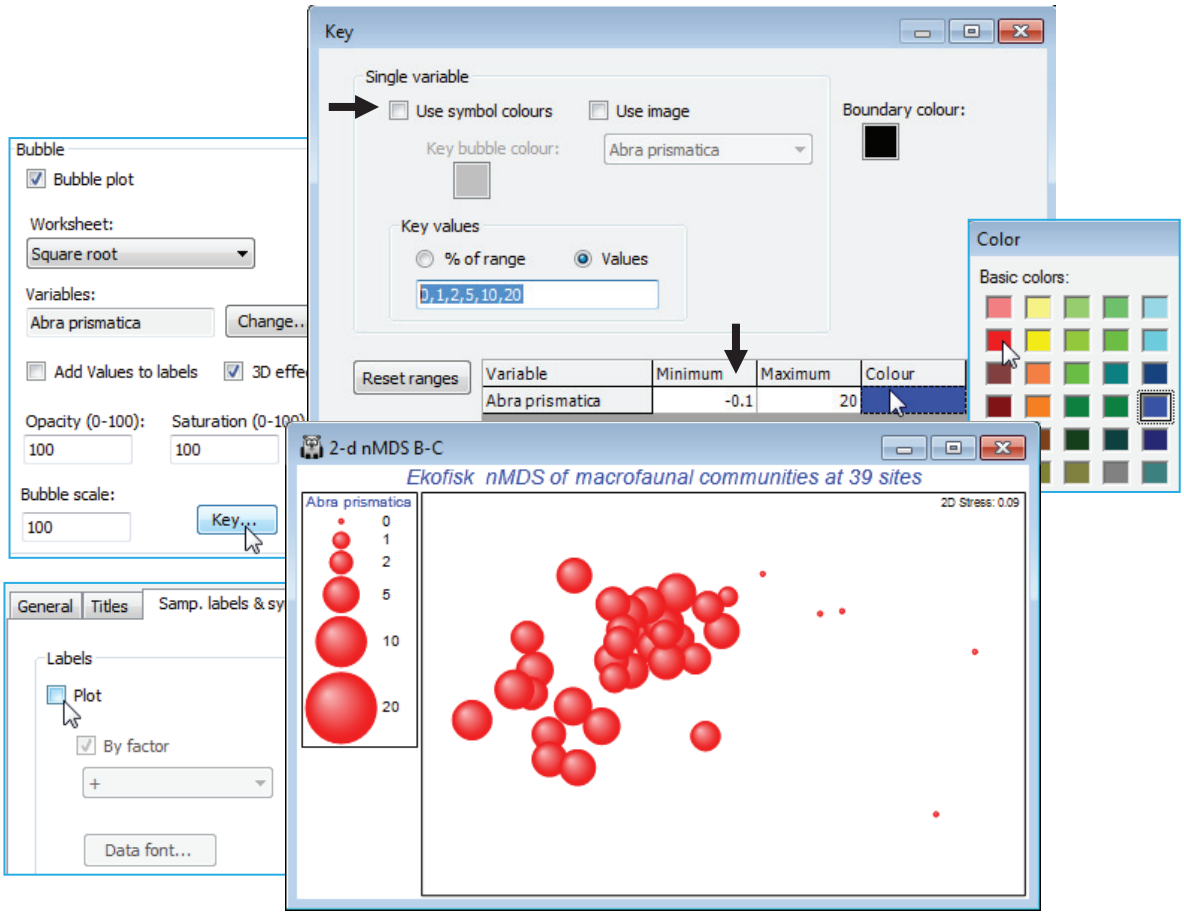
A new feature in PRIMER 7 is that these bubbles will be plotted in different colours – or in mono hatching patterns if the **General>**(✓Monochrome) option is selected – according to levels of the Symbol factor, in this case the four distance groups from the oilfield (A to D). At least, this is what will happen provided the **Key** dialog, accessed from the Bubble section of the **Special** dialog, has not been changed from the default of (✓Use symbol colours). A further feature is that colours can now be plotted opaquely (default) or with a continuous degree of transparency, and/or with bubbles scaled to a smaller maximum size, both of which may be advantageous here, where many bubbles are plotted in close proximity. These features are again controlled with the **Special>Main>**Bubble dialog, e.g. (Opacity: 75) & (Bubble scale: 70), leaving the colour saturation at (Saturation: 100). A further new feature is (✓3D effect), which comes into its own for 3-d ordinations (the default then).



Bubble key

v7 The key to the left of the plot, which gives the root-transformed scale for *Abra prismatica* counts, has been defined automatically and may not therefore be optimal – the values, and even the number of categories, are now (in PRIMER 7) user-selectable. The smallest non-zero value can only be 1 (the square root of 1) because the original data are integers, so a circle category of size 0.7 in the key is artificial, and not seen in the plot. We shall want to replicate this bubble plot ordination for several species, on a common size scale, to judge respective contributions to the observed *n*MDS pattern. Given that the maximum value in the transformed sheet, Square root, is somewhat over 20, for *Chaetozone setosa* – use **Analyse>Summary Stats>**(For•Variables) to see this – a natural key would have (square root) sizes 1, 2, 5, 10 and 20 (counts 1, 4, 25, 100 and 400). Take **Key** in the Bubble section of the **Graph>Special>Main** dialog box, run on the above 2-d nMDS B-C plot, change to (Maximum 20) and, under Key values, (•Values 1,2,5,10,20). The (Key bubble colour:) can be changed from its neutral grey by clicking on the grey box, to give the usual colour choice dialog, but this is not often likely to be advantageous because a less neutral colour would become confused with the colour codes for (symbol) groups A-D, if these are being employed (as above). However, given that the *a priori* grouping of these sites into distance ranges is a rather arbitrary categorisation of a likely continuous scale of impact and therefore response of the biota, e.g. with steadily increasing dilution of contaminants, it might be preferable to plot bubbles without this categorisation, by unchecking the (✓Use symbol colours) box on this **Key** dialog. The bubbles default to the standard blue symbol colour, but this can be changed in the (Colour) cell – this automatically changes the key to the same single colour. (The boundary colour for circles – only used when the 3D effect is not selected – can similarly be changed, from black, on this dialog).

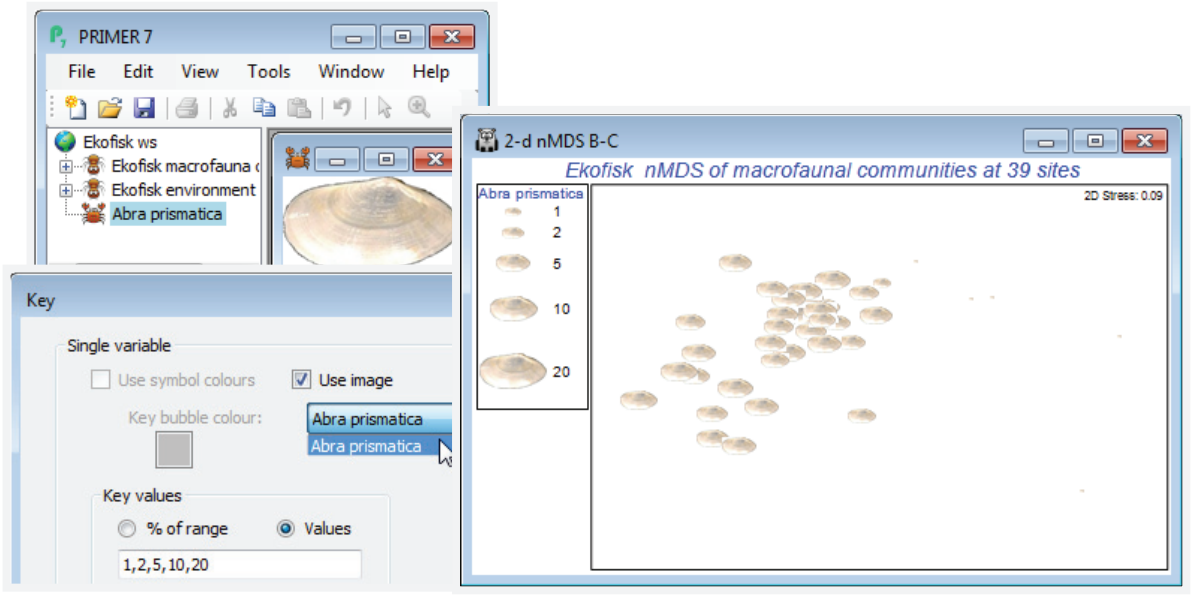
v7 Note that the plot below also changes back to the defaults of (Opacity: 100) & (Bubble scale: 100), from the **Special>Main** tab, and edits the title, removes the subtitle and the history box, from the **General** tab. And it employs a generally neater way of indicating the position of sites on the MDS where this species has a zero value. The '+' labels are removed, using the **Samp. labels & symbols** tab, and the Minimum cell on the Bubble **Key** dialog is set to a small negative number (here -0.1) instead of 0, ensuring that a tiny bubble is plotted at all sites. There is no risk of confusion in this case because there are no very low non-zero values, the lowest square-root count being 1, but this could be made doubly clear by adding zero to the Key values, i.e. using (•Values 0,1,2,5,10,20).



Bubble images

v7

A further new option on this **Key** dialog is to replace single-coloured circles with a (rectangular) supplied image plotted at different sizes. To illustrate this, within the C:\Examples v7\Ekofisk macrofauna directory there is a photo image of a single valve of *Abra prismatica* as an *.jpg file (at very low resolution in this case, though it could be much higher – avoid using an image of several Mb for this purpose, however, if you do not want to slow the MDS plotting noticeably). **File>Open** this, taking the file type as (Image file: *.jpg, *.jpeg, *.png, *.bmp, *.tif, *.gif, *.emf). Take **Key** in the **Special>Main>**Bubble section, and (✓Use image>Abra prismatica), which only operates when (✓Use symbol colours) is unchecked. Whether such a graph is successful depends on the image – the plotted shapes are rectangles not circles (and fully opaque, whatever the opacity setting) so it may be most effective for ordinations with relatively few points.



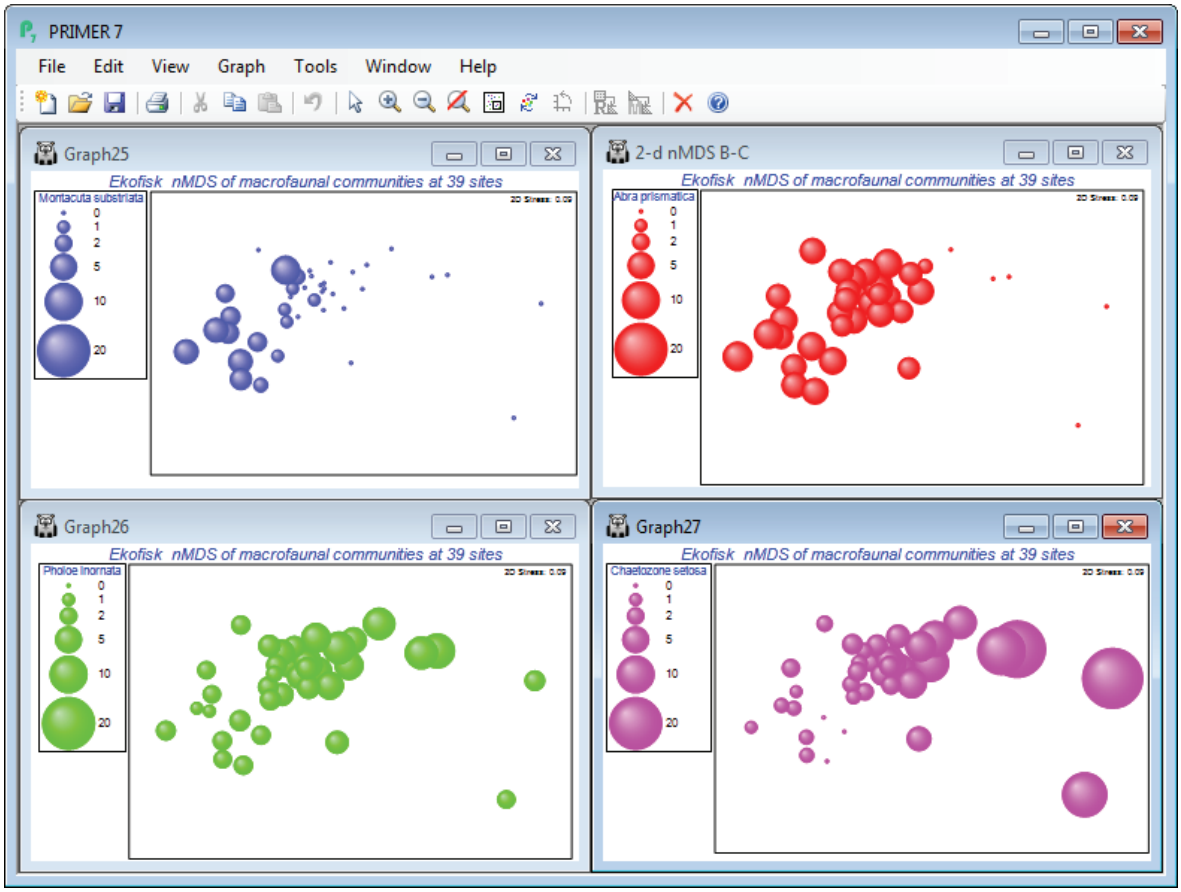
Duplicate graphs

Abra prismatica appears to be a species which can tolerate the contaminant or disturbance effects found relatively close to the oilfield since the plot shows that it is present in consistent numbers from sites 8km distant to those >250m – remember mentally to square up numbers in the bubble key's square root scale, or replot the ordination using the original matrix as the secondary data, (Bubble>Worksheet: Ekofisk macrofauna counts), rather than the Square root sheet. Either matrix is perfectly valid to use as bubble sizes but the interpretation is different. The above graph places the focus on how the species contributes to the ordination pattern, whereas superimposing raw data is a valid way of seeing how the actual counts vary from site to site. To examine a range of species contributions, firstly revert to the bubble plot at the bottom of the previous page – by unchecking (Use image) on the **Key** dialog – and take **Tools>Duplicate>**(•On existing branch) on this plot, three times. Then **Window>Close All Windows**, re-display the four plots by clicking on them in the Explorer tree, and **Window>Tile Vertical** – the example below also suppresses the Explorer tree using the **View** menu. (You could alternatively put all four plots into a new multiplot).

On the first copy, use **Special>Main>**Bubble>**Change**, taking *Montacuta substriata* over to the Include box and *Abra prismatica* back to Available. On **Key**, change its colour to blue (if needed), and again its Minimum to -0.1 and Maximum to 20. Note that if you change the variable plotted as bubbles, the routine automatically rescales to a value for the maximum bubble which is relevant for that new variable. We may often need this rescaling when plotting actual counts, since one or two species could be dominant and a universal abundance scale then reduces most species to invisible bubbles (the reason for transforms in multivariate analysis!). Separate scaling would also be needed for abiotic variables superimposed as bubbles on a community ordination, since there is then no common measurement scale – see later. But when the focus is on judging relative contributions different species make to the observed community gradient (as here), a common transformed scale is needed, so the min and max values of -0.1 and 20 need to be set for each copy of the MDS plot.

v7

(If you ever need to reset to default ranges for variables, use **Reset ranges** on the Key dialog). Now **Change** the second copy of the plot to *Pholoe inornata* and green bubbles, and the final copy to *Chaetozone setosa* and purple bubbles, and note the very different patterns across the gradient.



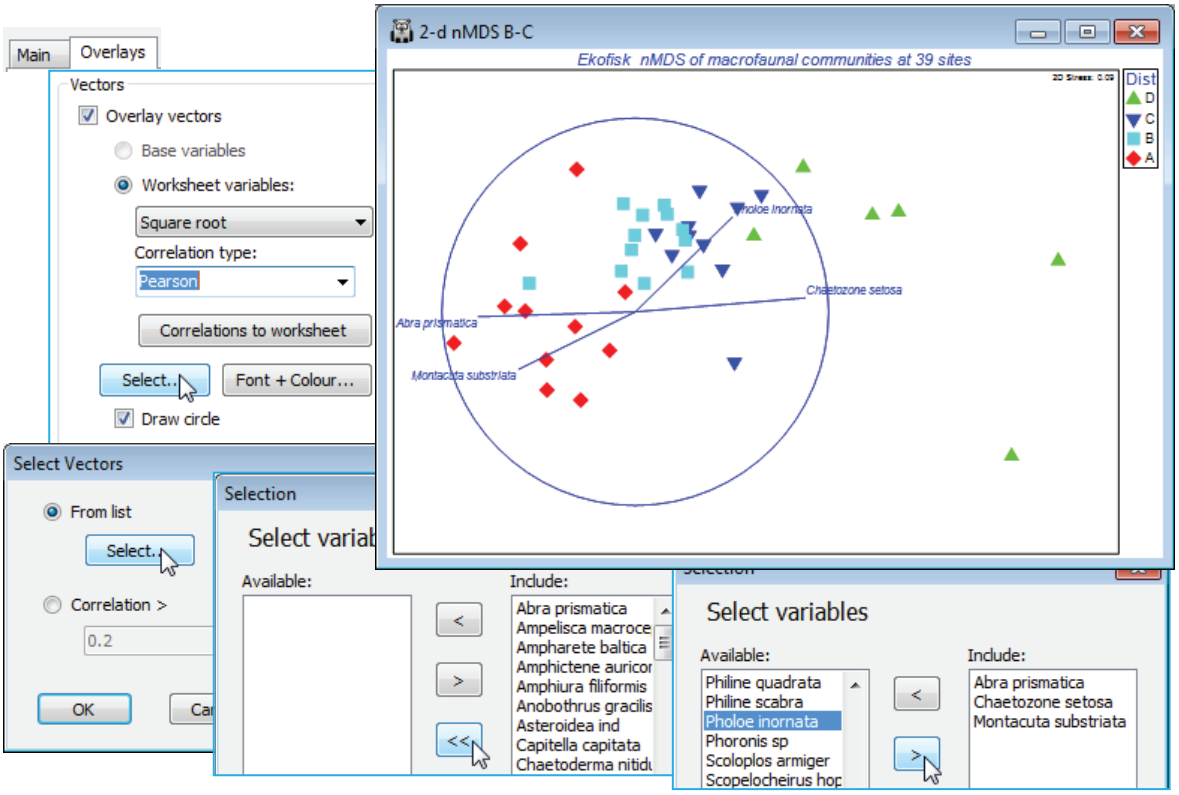
Bubble plots are usually clear and interpretation straightforward – however, there are 174 species here, many of which will show no pattern across the sites or will have such low numbers (seen in negligible-sized bubbles on this common scale) that they can contribute little to the site similarities. (An example is seen for these data in CiMC Fig. 7.13 for *Spiophanes bombyx*). Having established by hypothesis testing that there are meaningful community patterns (e.g. ANOSIM, Section 9, on the distance groups, or RELATE, Section 14), to justify interpreting species patterns, one can avoid having to plot all 174 bubble plots(!) by examining Shade Plots and SIMPER results (Section 10), to pick out species for which bubble plots will be worthwhile.

v7

Vector plots for species

One tempting alternative, commonly found in the literature, is to plot (species) variables as vectors on the MDS plot (see below). They are directional lines emanating from a common origin, pointing in the directions in which those species numbers increase, and of length reflecting the strength of pattern in those species counts along that direction. Such a vector plot can have its origin anywhere (there is a common misconception that its placement or overall size matters) and PRIMER chooses to locate it the left side of the plot – deliberately non-central to make the point about arbitrariness of its origin. The perceived advantage of a vector plot is that it can display the relationship of many species to the ordination configuration in a single plot. But, in reality, it rarely succeeds. Firstly, a vector assumes the relationship is strictly linear, but the above bubble plots for *Abra prismatica* and *Montacuta substriata* show threshold relationships, present at consistent numbers until a (different) point is reached, closer to the oilfield, when numbers drop to zero. Even worse is *Pholoe inornata* whose relationship to the gradient is not even monotonic, because it is an opportunist increasing in numbers closer to the impact but decreasing again when the impact becomes strong, within 100m. Of the above four bubble plots, only *Chaetozone setosa* would be well described by a linear fit of (root) counts to (*x*,*y*) co-ordinates of the MDS plot. Secondly, vector length measures (Pearson) *correlation* of transformed counts to these axes (see Chapter 7 of CiMC), not just a purely linear construction but also one that ignores the universal scaling of the bubble plots – if a species has very low numbers throughout and thus could not contribute much to the similarities and ordination, it might still have a large vector if its numbers gently decrease (or increase) closer to the oilfield. In other words, vector plots only make sense in the bubble plot context above, of judging relative contributions to the ordination, where species have been standardised prior to analysis, i.e. where species are given the same weight in similarity calculations, irrespective of their overall abundance.

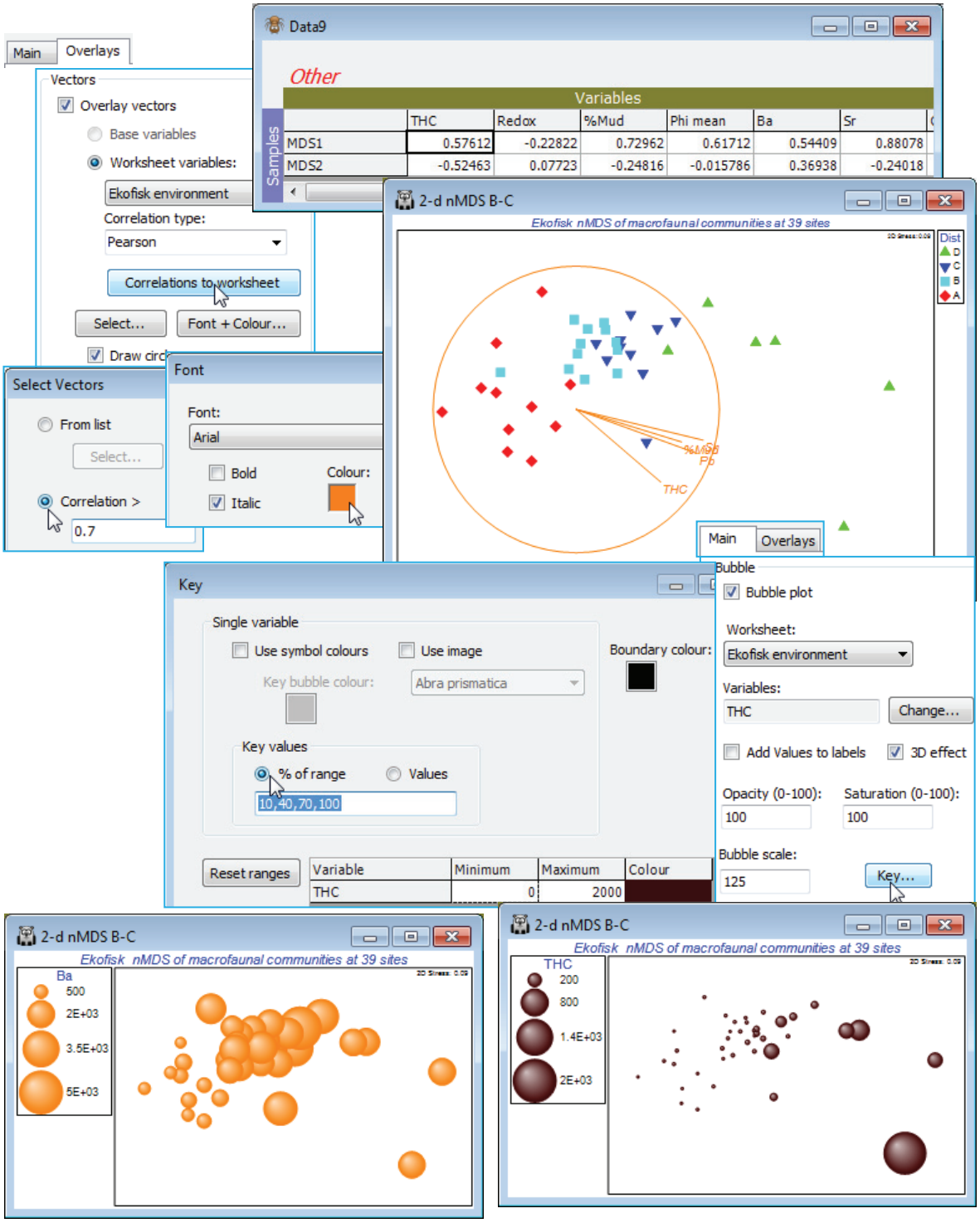
These are compelling reasons for never using a vector plot of species! Nonetheless, a vector plot can be implemented as follows. On the MDS graph 2-d nMDS B-C, under **Graph>Special** turn off (✓Bubble plot) on the **Main** tab, and on **Overlays** take (Vectors✓Overlay vectors)>(•Worksheet variables: Square root)&(Correlation type: Pearson)&(✓Draw circle)&(**Select>**•From list>**Select**), and leave only species *Abra prismatica*, *Chaetozone setosa*, *Montacuta substriata* and *Pholoe inornata* in the Include box. The circle indicates a (multiple) correlation of 1 of that species to the 2-d MDS *x*, *y* co-ordinates (see the 'Vector plots' heading in Chapter 7 of CiMC for discussion of this and the alternative(s) to Pearson correlation). Clearly, the resulting vector plot carries limited information compared with bubble plots: the different threshold responses of *Abra prismatica* and *Montacuta substriata* cannot be captured, likewise the non-monotonic relation for *Pholoe inornata*.



Environment bubble & vector plots

An important use for a bubble plot is in displaying the behaviour of a further measured variable, which has not been used in the multivariate analysis, across the sample positions on an ordination, e.g. in superimposing environmental or contaminant variables one at a time, on a community MDS. Also, in this case, vector plots have a more useful role because there is a better chance of linearity, or at least monotonicity, of the relationships between abiotic gradients and community gradients – also the rescaling of abiotic values implicit in a correlation coefficient is appropriate here, because environmental variables will usually be on different measurement scales.

As an example, **File>Open** the Ekofisk environment sheet into the current workspace (if it is not already there from the transformation examples of Section 4). On the biotic 2-d nMDS B-C plot, under **Special>Overlays**, retain (✓Overlay vectors) but change to (•Worksheet variables: Ekofisk environment) & **Select>**(•Correlation > 0.7) to display only those abiotic variables with a high (multiple) correlation coefficient with the biotic ordination positions. The vector plot below shows these to be *THC, %Mud, Sr* and *Pb*. Those multiple correlations can be calculated by clicking the **Correlations to worksheet** button on the (✓Overlay vectors) dialog, which produces a worksheet of Pearson correlations of each abiotic variable with the ordination co-ordinates (MDS1, MDS2). These are the values obtained by projecting each vector onto the (*x, y*) ordination axes, which is how the vectors are constructed in this case. The length of the vector (the multiple correlation) is simply obtained by Pythagoras, e.g. for *THC*, $\sqrt{\{0.576^2 + (-0.525)^2\}} = 0.78$. [For the technically minded, that this is equivalent to the multiple correlation coefficient from multiple linear regression of *THC* on the (MDS1, MDS2) co-ordinates follows from the fact that the MDS solution is rotated to PC's, making the MDS axes uncorrelated]. In addition, simple bubble plots (e.g. of *THC* and *Ba*) can be drawn as before, using **Main>**(✓Bubble plot>Worksheet: Ekofisk environment).



The vector plot, though not as unsatisfactory as when superimposing species, can again be seen to miss an important relationship here by comparison with the bubble plots. The pattern of *Ba* is not a linear one and thus does not give a very strong correlation, but it is distinctive and instructive, since barite is a component of the drilling muds dispersing into the marine environment. The bubble plot shows that *Ba* values are consistently higher around the oilfield but drop to background levels at about 3-4km – another example of a threshold change not optimally captured by correlation. Note that plots here used untransformed abiotic variables, which is visually informative, but there is a good case for instead using the selectively transformed sheet (Data3) deemed appropriate for these variables in Section 4; e.g. this would reduce the dominant effect of the large outlier in *THC*.

Segmented bubble plots

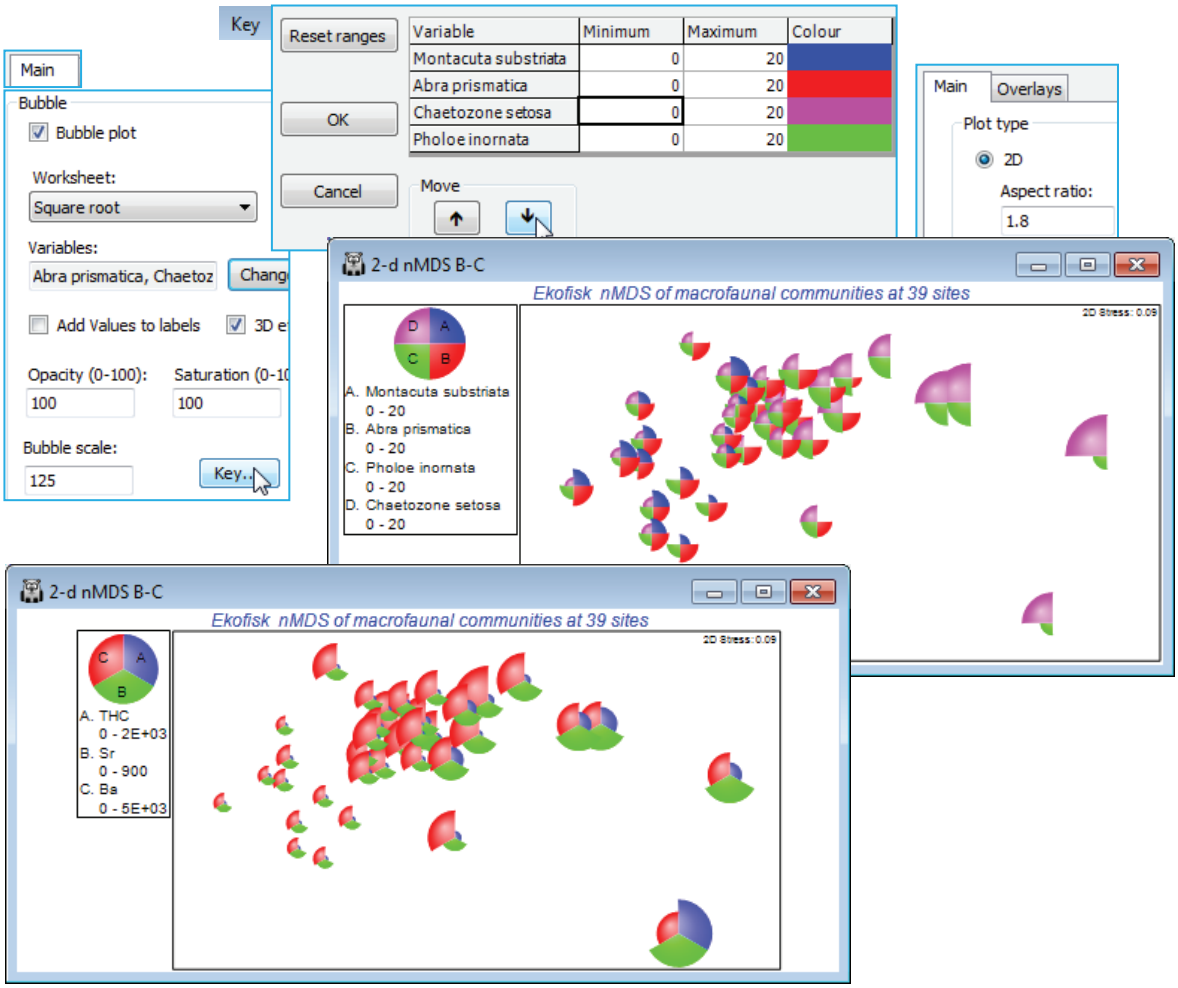
v7

Given the inadequacies of vector plots for displaying species contributions to ordinations, but the practical need to reduce the number of graphs which can be displayed in publications, PRIMER 7 introduces a *segmented bubble plot*, which attempts simultaneously to display the bubble sizes for several species (or abiotic variables) on the same ordination plot, technically as circle *sectors* (or

v7 their spherical equivalent in 3-d), with a species always occupying the same sector position but at differing sizes. Careful choice is needed of variables to juxtapose, their colours and positions, and whether common or individual scaling is used, but some of the same flexibility is available as for single-variable bubble plots (options not offered here include user-supplied images and the ability to represent different factor levels by different colours – that would be too confusing by far!)

Returning to the display of individual species counts (square-rooted) on the Ekofisk 2-d nMDS B-C ordination, the four separate bubble plots seen earlier can be combined into a single plot by taking v7 **Graph>Special>Main>✓**Bubble plot>(Worksheet: Square root)>**Change** and moving *Chaetozone setosa, Montacuta substriata* and *Pholoe inornata* to join *Abra prismatica* in the Include box. **Key** now lists all four species in the Key dialog, with their differing default ranges (e.g. 0 to 7 for *Abra prismatica* but 0 to 30 for *Chaetozone setosa).* As before, these need to be harmonised (e.g. set to 0 to 20) so that the relative contributions each of these species is likely to make to defining parts of the community gradient can be properly assessed. Also change the colours to match those for each species used on the single plots (three pages ago) and, since there is a natural progression of species (not alphabetic!) in terms of where on the gradient they are mainly found, use the Move up and down arrows to place them in the order: *M. substriata, A. prismatica, P. inornata* and *C. setosa*, which then become the sectors (quadrants in this case) reading clockwise from top right. Again, depending on how many points there are on the ordination (and this one is possibly too cluttered for a really effective segmented bubble plot) you might need to adjust the overall size of bubble plotted for the specified maximum count (square-rooted) of 20. The plot below uses (Bubble scale: 125) and, of course, the bubble size in the key will correctly reflect such a change. Another minor v7 detail is that, for this rather long thin ordination plot, the shape of the boundary has been changed to match that, with **Main>**Plot type•2D>(Aspect ratio: 1.8) rather than the default of 1.5.

A segmented bubble plot for some of the abiotic variables (*THC, Ba, Sr*), on this same ordination, is obtained in the same way, substituting (Worksheet: Ekofisk environmental) for Square root, the only change being that default ranges for each variable are used, since measurement scales differ. (Re-)save the Ekofisk workspace as Ekofisk ws, and close it.



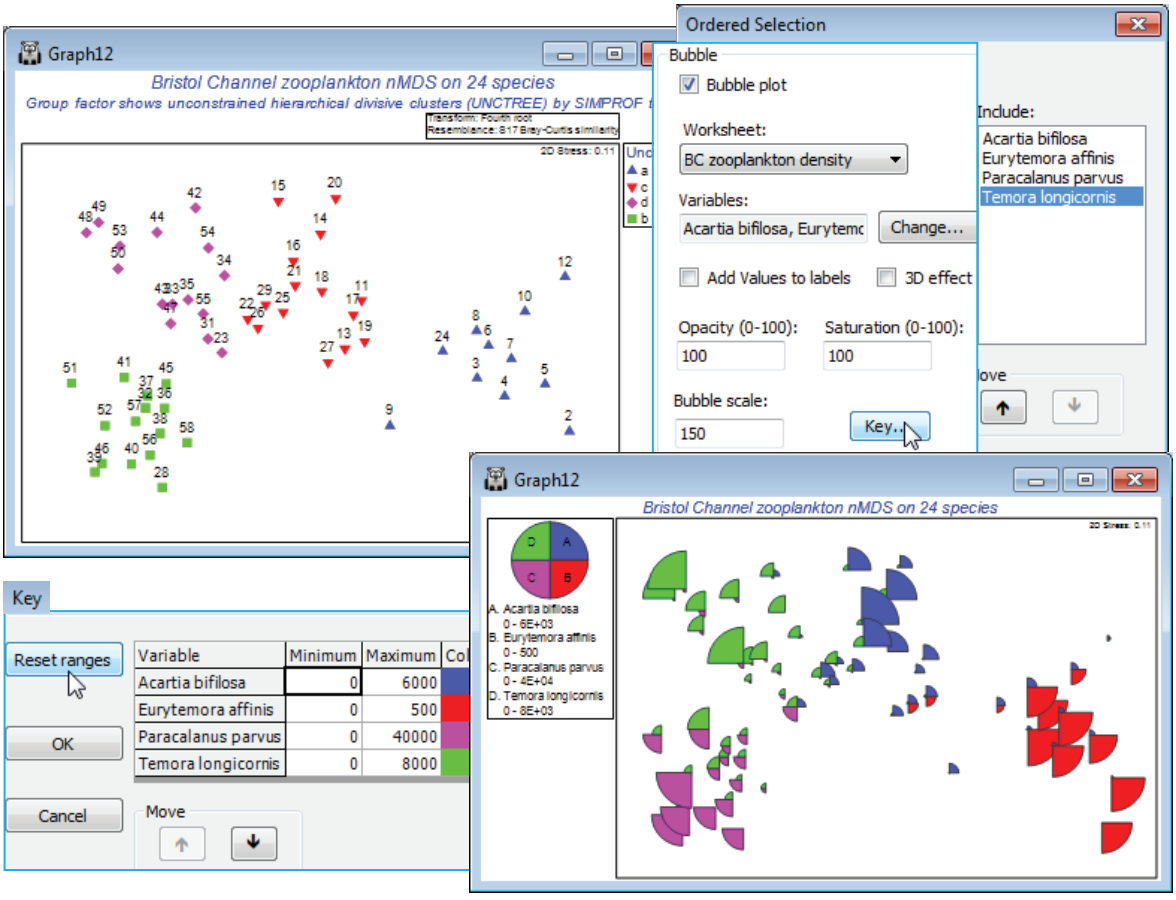
Segmented bubble plots work best either where species values change markedly over the points of the ordination, or there are relatively few points, such as in an ordination means plot. Examples of both these situations are found at the end of Chapter 7 of CiMC, on data sets that we have already met, so you may like to reproduce them, as below.

(Bristol Channel zooplankton)

v7

The BC zooplankton density data sheet in C:\Examples v7\BC zooplankton was used in the last half of Section 6 to illustrate different clustering methods, and the workspace Bristol Channel ws may be available from that. If not, open the data, fourth-root transform it, compute Bray-Curtis similarities and construct an *n*MDS plot. SIMPROF tests in Section 6 produced 4 groups, broadly similar for all clustering methods, and these look convincing in the *n*MDS below, and in Fig. 3.10 of CiMC, though it is equally clear that they form part of a continuum of change. A shade plot and SIMPER analysis (Section 10) identify four species which most *typify* the clusters (*Acartia bifilosa, Eurytemora affinis, Paracalanus parvus* and *Temora longicornis*) and are good *discriminators* between them – these are shown on a segmented bubble plot below. The focus here is simply on observing how the actual densities for these key species vary over the site gradient (this is largely salinity driven). So, whilst the *n*MDS is computed on heavily transformed scales, the raw densities themselves are plotted as the bubble segment sizes, requiring different scaling, varying over an order of magnitude (or nearly two). This is in contrast to the previous example, where the focus was on identifying species actually contributing to the *n*MDS pattern, and for which common scaling of the transformed counts was then more appropriate.

v7



Bubble plots in 3-d MDS

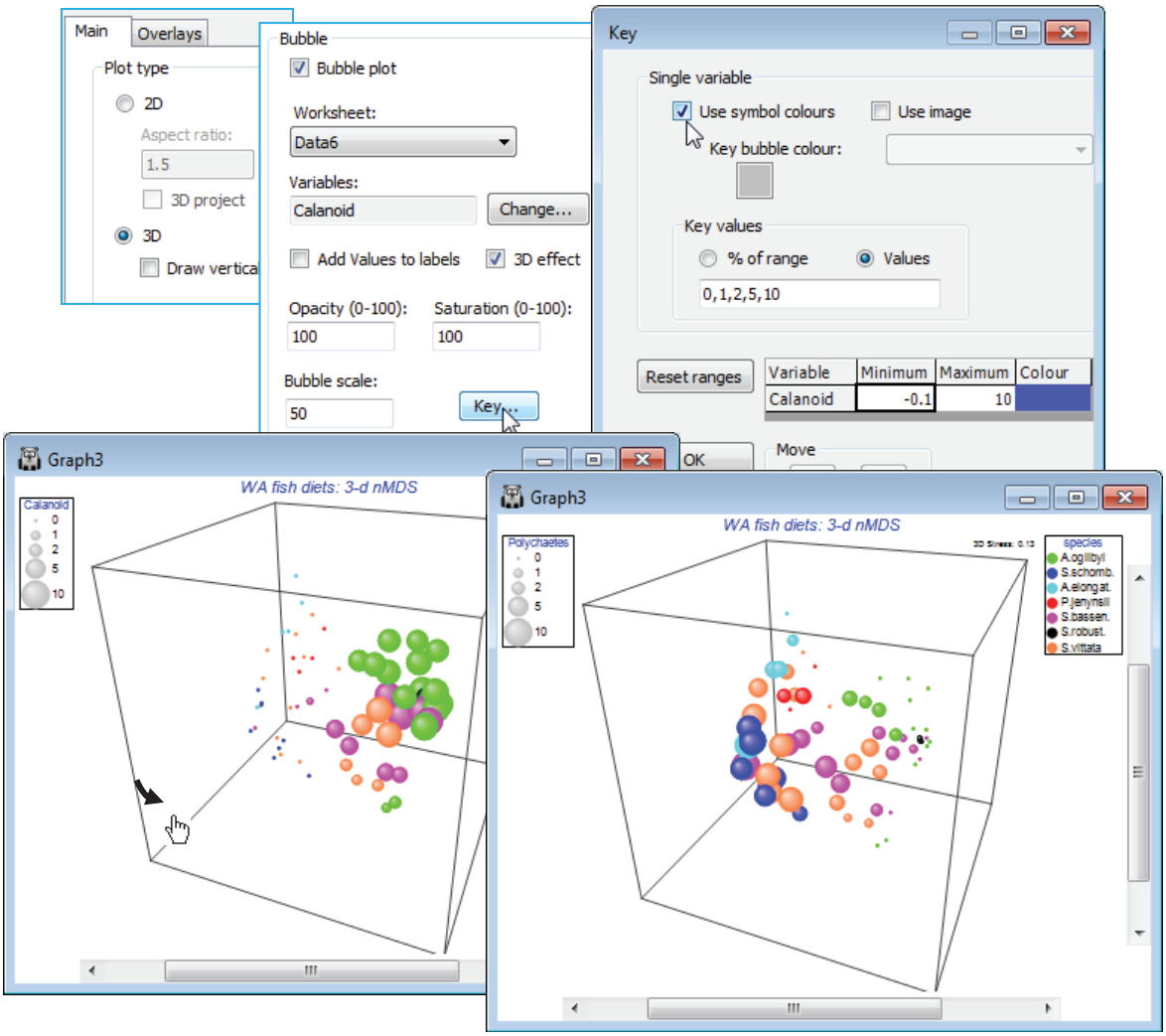
v7

Another new feature in PRIMER 7 is the ability to construct bubble plots in 3-d ordinations, where the same '3-d effect' representation of bubbles is employed as was seen earlier in 2-d plots – only now, of course, (✓3D effect) is the default. These are reasonably convincing as whole spheres, when the 3-d plot is rotated (as described earlier in this section), though inevitably less so as segmented bubble plots – though the latter plots are also permitted.

(W Australia fish diets)

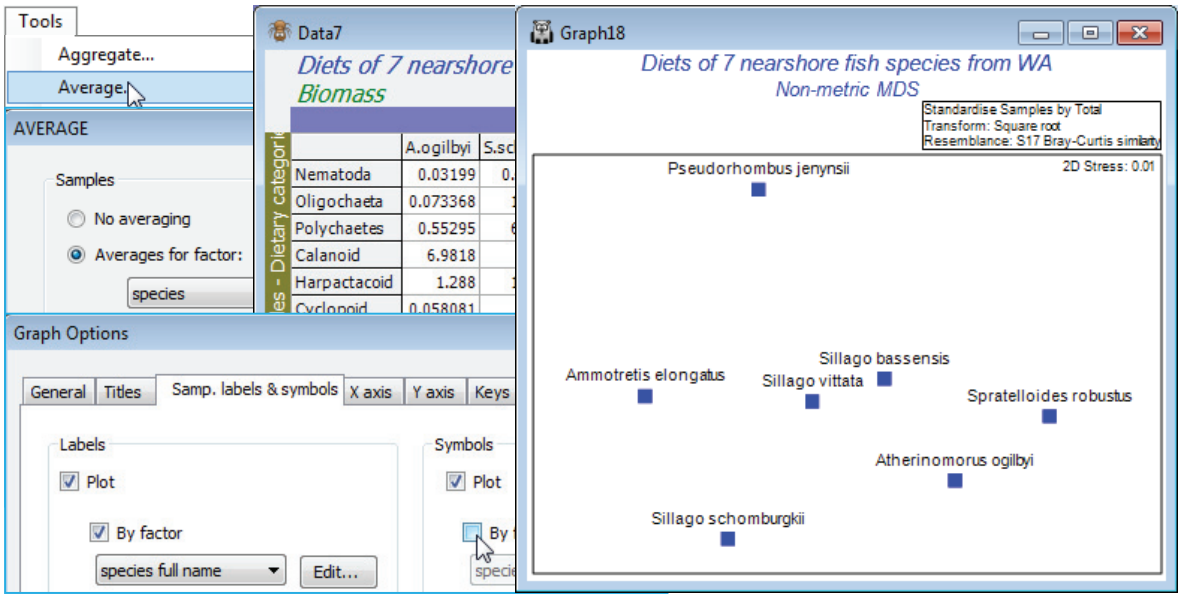
v7

To demonstrate bubble plots in a case where a 3-d ordination was necessary, in order to reduce stress to acceptable levels, re-open the WA fish ws workspace used earlier in this section (to show spinning and digital recording of 3-d solutions). On the 3-d *n*MDS plot, Graph3, create individual bubble plots for the dietary constituents *Calanoid* and *Polychaetes*, using the secondary data for which samples have been standardised and then square root transformed (probably Data6).



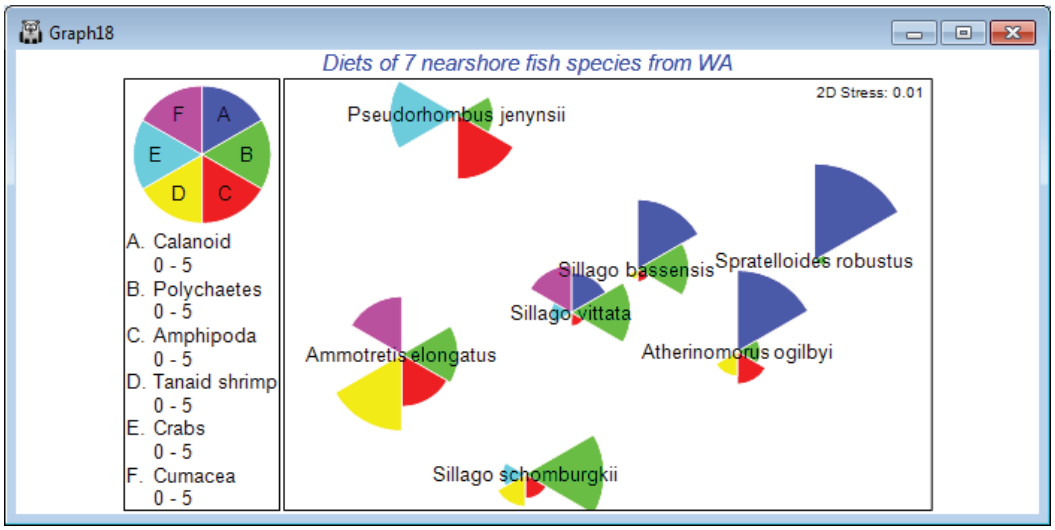
Bubble plots on averages

For these data, a segmented bubble plot on a 2-d ordination provides a succinct summary of the relative balance of the main dietary components for these 7 fish species, when examined on data averaged over all replicate gut samples for each fish species (as in Fig. 7.16 of CiMC). On Data6, the standardised then transformed data sheet, take **Tools>Average>**(Samples•Averages for factor: species), giving a matrix of 7 samples (the 7 fish species) by 32 dietary categories, Data7. Compute Bray-Curtis similarities on this and produce the 2-d *n*MDS configuration, which has minimal stress and is therefore an excellent representation of the dietary dissimilarities among the fish species. The *means plot* below labels by the full species name and harmonises the symbols to a blue square.



v7

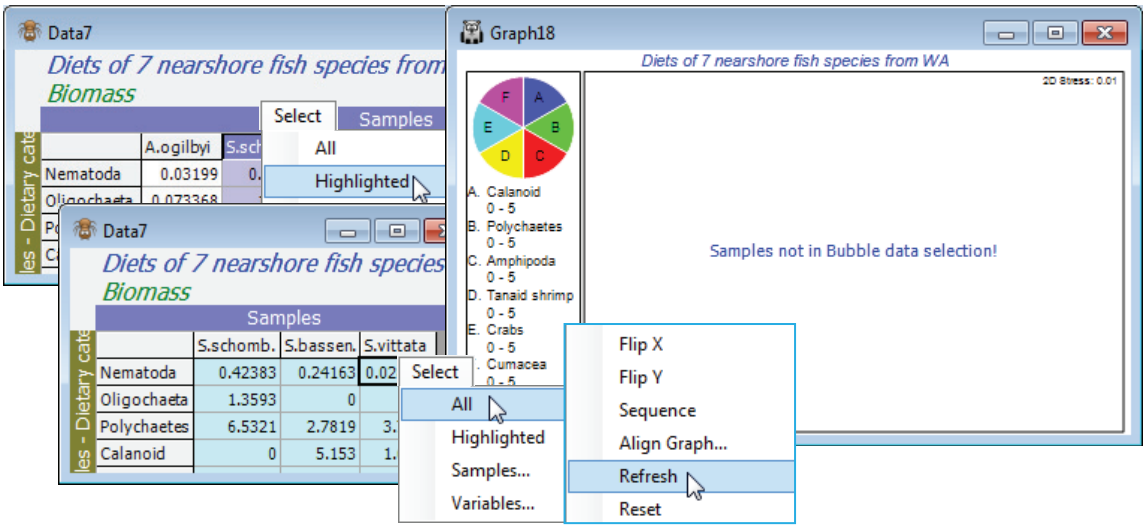
Now, uncheck the History box on the **General** menu, remove the subtitle from the **Titles** tab and turn this into a segmented bubble plot by **Special>Main>**✓Bubble plot>(Worksheet: Data7), and **Change** variables to *Calanoid (copepods), Polychaetes, Amphipods, Tanaids, Crabs, Cumaceans*, also increasing the Bubble scale. On the **Key** dialog, set a uniform range of 0-5 (square root scale, so the segment sizes in the key will represent an average 25% of diet by volume), change Boundary colour to white and make any other desired colour changes. (If the names are required displaced from the bubble centres, or the key back-transformed to original units, key/titles moved etc, as in the equivalent Fig. 7.16 of CiMC for example, the plot can be copied and pasted to Powerpoint or other presentation software; this will be in vector not bitmap format so can readily be edited). Resave this fish diets workspace as WA fish ws for use in the next section, and close it.



Bubble data selection error & Refresh

It is worth repeating the point made at the start of these examples on bubble plots: if the routine cannot find one or more of the sample labels of the MDS points in the supplied secondary data sheet of bubble values, you will get a blank graph with the message *Samples not in Bubble data selection!*, or if it cannot find the worksheet at all, *Bubble data worksheet is missing!* These are quite common error messages. The latter may arise because, while the secondary data sheet was correctly supplied and used to create the bubble plot initially, it was subsequently deleted or, more likely, had a name change. You will not identify this until the bubble plot window is shut down, perhaps as a result of saving and closing the workspace. When it is reopened, the MDS plot will display the error because it can no longer find the linked sheet name it was initially given. More subtly, the *Samples not in Bubble data selection* can occur because the secondary data sheet has, subsequent to creating the bubble plot, been subjected to a selection of its samples (for some other purpose), and left in that state, so that some of the MDS points cannot be matched in the current selection. If this is the problem, then **Select>All** on the secondary sheet will solve it, but note that when returning to the MDS plot, the error message is only removed by either shutting the window down (with) and reopening it or, more neatly, by **Graph>Refresh**.

v7



A general point to note is that though all the overlays given above (trajectories, bubble, vector plots etc) have been applied to *n*MDS plots, with only minor exceptions they are generic and available in any ordination configuration plot, e.g. PCA (Section 12), PCO and CAP (in the PERMANOVA+ add-on to PRIMER) and metric MDS (*m*MDS), which we now describe.

Metric MDS

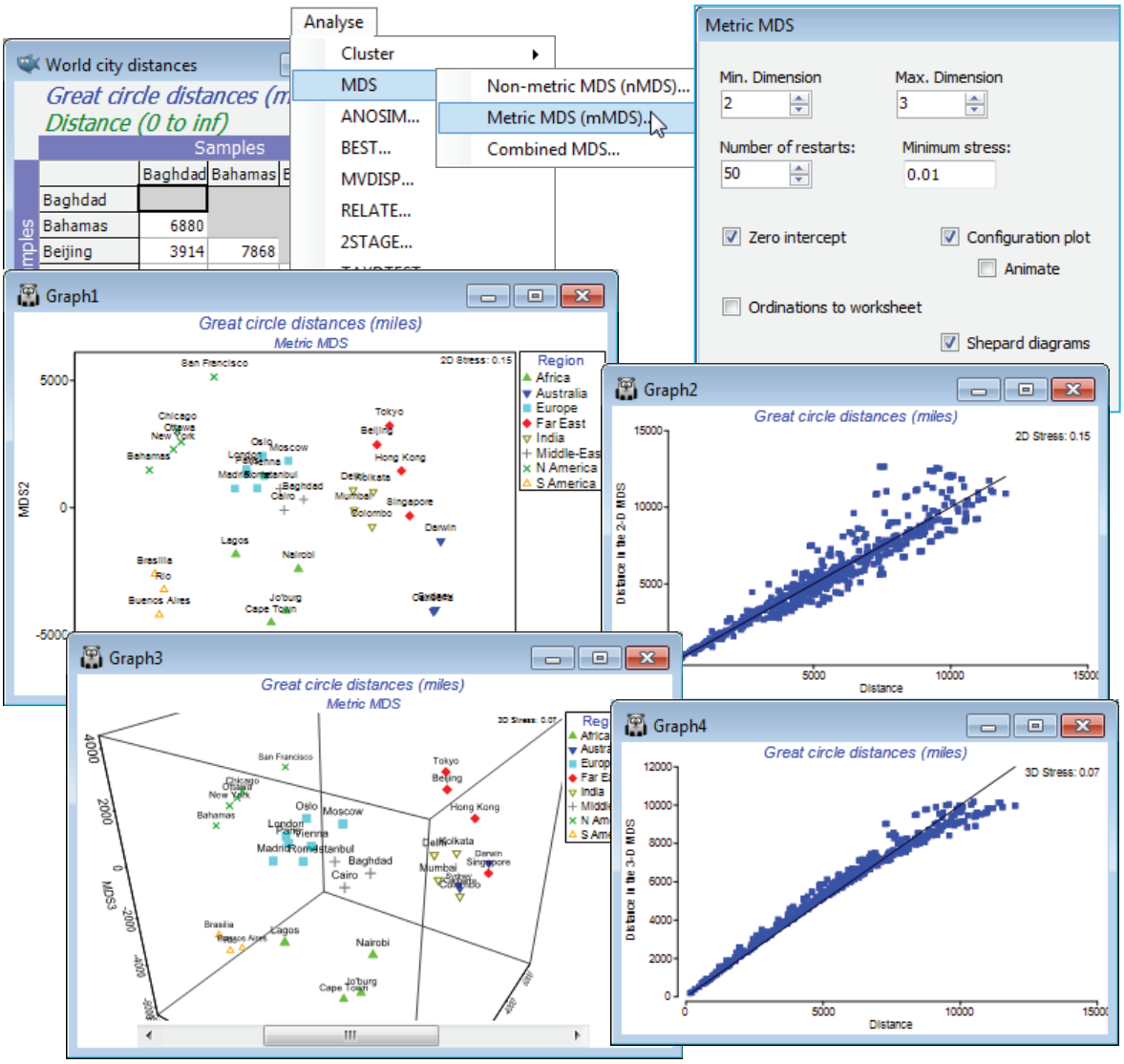
v7

A significant new feature in PRIMER 7 is the construction of *metric MDS* ordinations (*m*MDS), together with variations such as *threshold mMDS (tm*MDS*)* and a specific combination of *n*MDS and *m*MDS; the rationale and details are fully covered towards the end of Chapter 5 in CiMC. The *m*MDS algorithm operates in a very similar way to *n*MDS, by an iterative search for an optimal configuration of samples in a specified dimension (or, in practice, a range of dimensions). The key difference is in the Shepard diagram, which is constrained to be fitted by a straight line through the origin (i.e. two points should only be represented as coincident if their dissimilarity is 0), as a result of which *m*MDS has axis measurement scales. Under conditions of low stress and a well-fitting line in the Shepard diagram, the distances in the plot are thus directly interpretable as the dissimilarities/ distances in the original resemblance matrix. This is only likely to happen, however, either for a resemblance matrix that is close to behaving like Euclidean distances in the first place, e.g. from an environmental analysis using normalised Euclidean distance, or where there are only a handful of samples to ordinate, e.g. for some means plots. In such cases, *m*MDS has significant advantages over *n*MDS; in normal community analyses *n*MDS will usually be much superior.

(Great-circle distances for world cities)

v7

A data set for which the resemblance matrix appears unequivocally suited for *m*MDS is discussed in CiMC Chapter 5 and given in directory C:\Examples v7\World cities, data World city distances. This is a physical distance matrix in the first place, being a triangular matrix of the (great-circle) distances round the globe between all pairs of 33 world cities (drawn from The Reader's Digest Great World Atlas, 1962). Run **Analyse>MDS>Metric MDS (mMDS)** under the defaults, to give both a 2-d and 3-d map of these (3-d) cities, along with their associated Shepard diagrams.

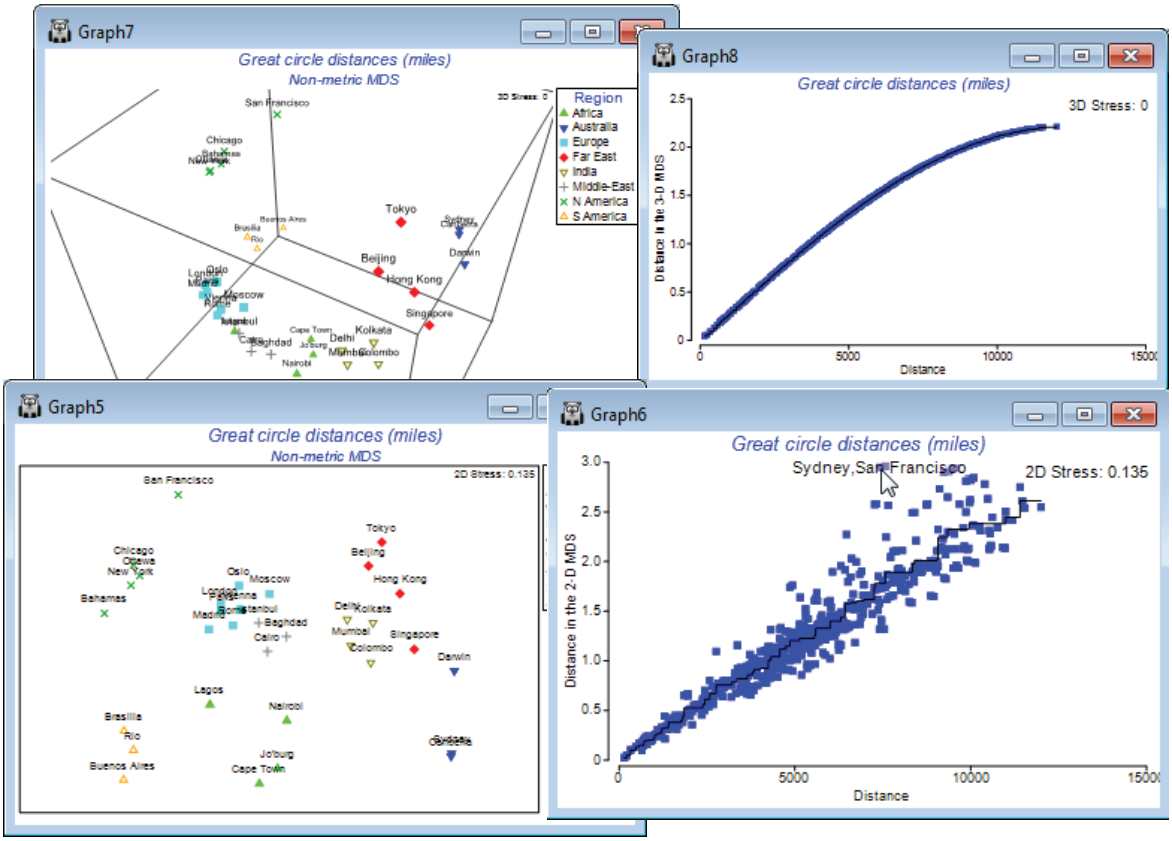


You will probably need to **Graph>Flip X** or **Y** and **Rotate Data** to get the standard 2-d orientation of world maps, and also **Rotate Axes** for the 3-d plot, though the arbitrariness of compass points (N at the top etc) is clear! Note how the axis scaling on the *m*MDS plot subtly changes as the data is rotated within the plot, as it needs to do in fully preserving the meaning of distance. Though the 3-d *m*MDS does place the cities in roughly the correct relation to each on the surface of a sphere, it may initially be a surprise to see that this 3-d picture of 3-d distances has non-negligible stress (of 0.07). The clue is in the Shepard diagram, which shows that the data is trying to force the relation away from the fitted straight line of the metric solution towards a curve, especially for the larger distances. Remembering that the matrix is of *great-circle* distances (direct air miles) but the 3-d MDS represents *through the Earth* distances resolves the issue. The true relationship is not linear and, even for this (apparently simple) physical distance matrix, *n*MDS is superior, placing the cities perfectly correctly on the globe surface, with stress 0, and marginally reducing the 2-d stress.

Identifying points on the Shepard plot

v7

Show this by running the matching *n*MDS to the above *m*MDS, this time setting (Minimum stress: 0.001). For the 2-d *n*MDS, identify the greatest contributions to the stress by clicking on outlying points in the Shepard diagram, which then lists the pairs of cities involved, straddled across the point. You can increase the font size of this text, if needed, by e.g. **General>Info font>**(Size: 120), – this is the button that also controls font size for the stress value shown here and on MDS plots. It is clear that several outliers involve distances to San Francisco, with the MDS attempting to move it closer to Sydney and Tokyo, in keeping with the true distances, whilst maintaining its distance from Europe and Africa etc – an impossible 'circle to square' in 2-d, hence the stress of 0.135.



Animating the *m*MDS/*n*MDS iterations

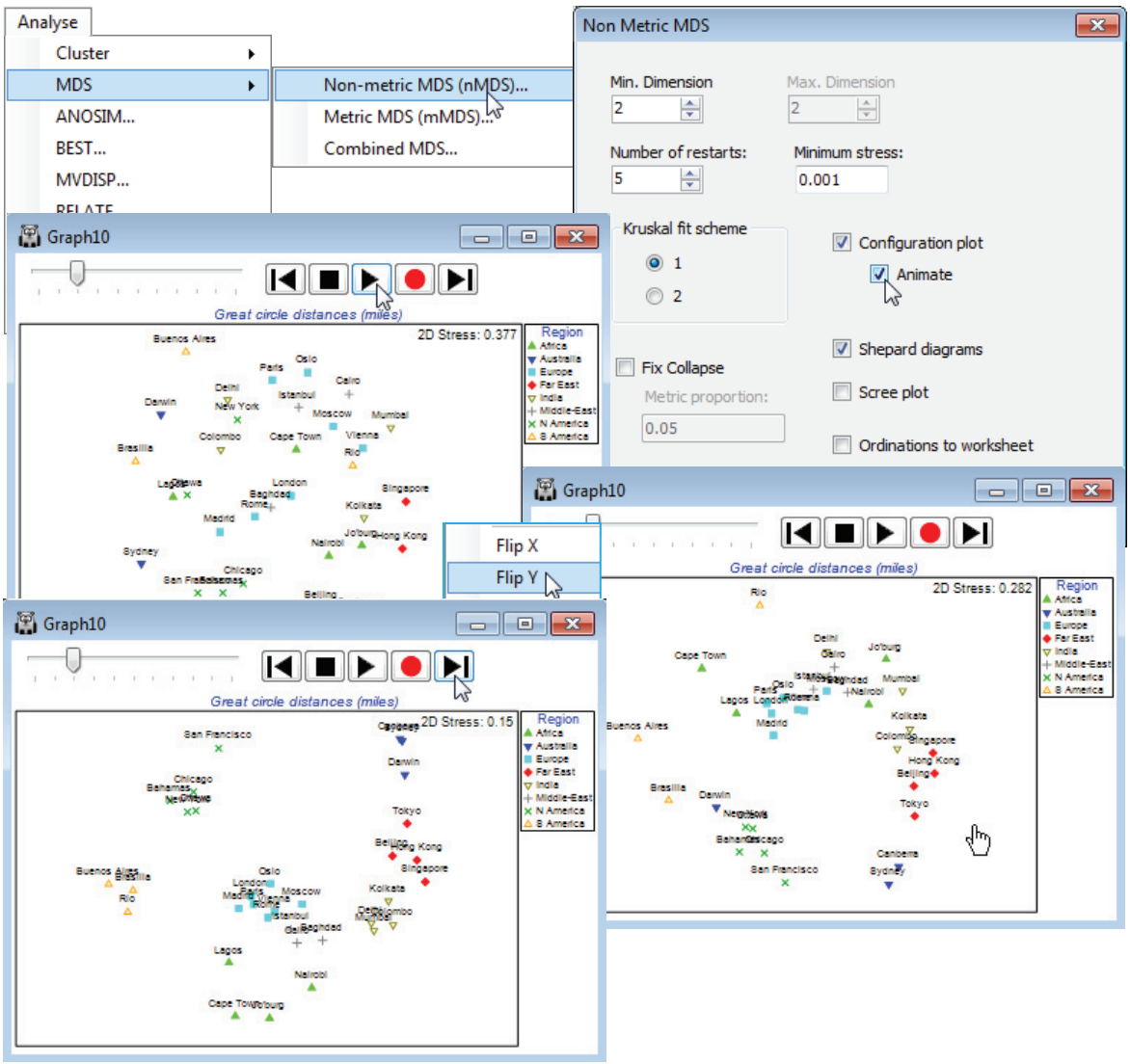
v7

Another feature introduced in PRIMER 7 is more of a teaching tool than an analysis method *per se*, namely the ability to watch the steps of the MDS iterative process (for either *n*MDS or *m*MDS) in converging – or failing to converge! – to an optimal 2- or 3-d representation of the relationships in the resemblance matrix. The option is taken by (✓Configuration plot)>(✓Animate) on the usual MDS dialog. It automatically greys out the (Max. dimension) box so that a single dimension needs to be chosen, via the (Min. dimension) box – for observation, 2 or 3 are the only realistic choices. It is wise, purely for this demonstration process, to reduce to (Number of restarts: 5), say, and to set (Minimum stress: 0.001) so that incremental improvements in stress during the iteration can be seen. On taking **OK**, the usual animation controls are displayed (as seen earlier in this section for spinning 3-d plots and sequence animations), and the first MDS iteration is started with ▶

A 2-d *n*MDS or *m*MDS of the World city distances provides a useful example for demonstrating the MDS iterative process because of the prior understanding of what a good 2-d solution would look like, so that an iteration which is heading into the cul-de-sac of a local minimum (see Chapter 5 of CiMC) will readily be appreciated. To improve the visual cues, a Region factor is defined, dividing the cities into continental areas with differing coloured symbols. A typical local minimum for this data set would, for example, be when the initial random starting configuration and the early iterative steps separate the African cities into two groups, either side of the denser European-Asian band of points. The routine may then have trouble re-uniting these two groups because any attempt to move African cities through the Europe-Asia axis, by small incremental steps, will be rejected because initially a move in that direction will always increase the stress. In this example, at least 50% of the restarts end up trapped in a local minimum, i.e. converging to a higher stress solution than can be found. Observing such local minima imparts a good appreciation of the importance of increasing the number of restarts routinely run, especially for complex MDS plots.

v7

The speed with which the iterative cycle takes place can be controlled by the slider, and the pause button ⏸ is usually necessary to re-orient or flip the axes to a recognisable layout after a grouping of the continents starts to emerge. Initial random configurations have stress values of around 0.4 and occasionally a restart will fail to get going at all but, more typically, stress falls quickly and then slows nearing convergence. At any point, the process can be paused with ⏸ and then stepped through incrementally using either of the end controls, ⏭ or ⏮. Use of the stop button, ⏹, is a premature cancellation of the whole run, returning no results. As for previous animations, recording can be started at any point by clicking the record button, ⏺ , which pauses iteration while details of picture resolution, frames per second and output file name (*.mp4 or animated *.gif) are given. The iteration and recording then restart automatically, and are terminated by ⏹ – in fact recording must always be terminated by this stop control (so ask for more restarts than you plan to record).



(Morlaix macrofauna, Amoco-Cadiz oil spill)

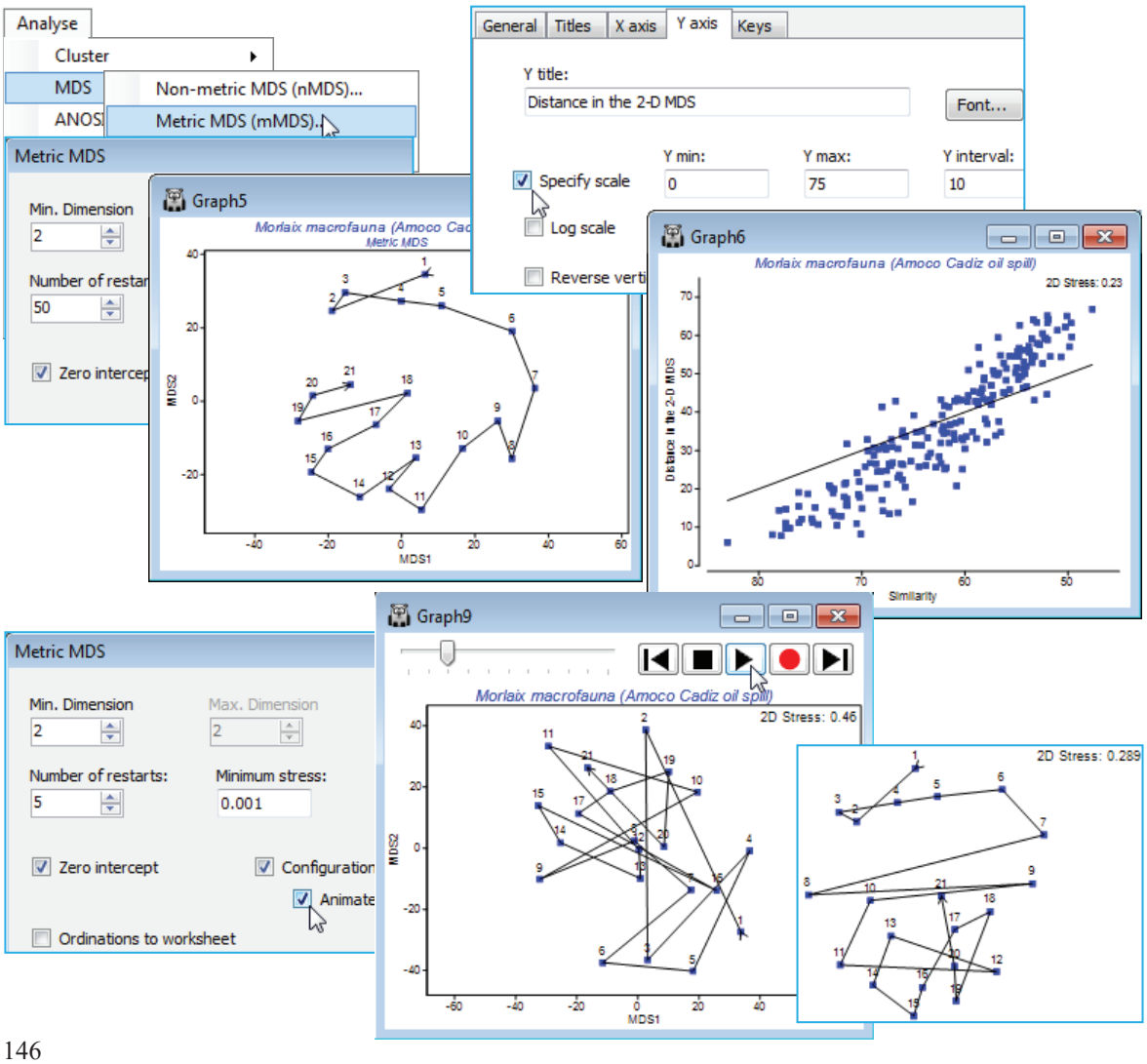
Close the World cities worksheet (it will not be needed again) and re-open workspace Morlaix ws, from earlier in this section. It contains the data sheet Morlaix macrofauna abundance, fourth-root transformed and with Bray-Curtis similarity matrix, Resem1, on which to carry out 2-d *m*MDS, with **Analyse>MDS>Metric MDS (mMDS)**, taking all the default options, and making the time factor the labels rather than the symbols, on **Samp. labels & symbols** – you should also join the points in *time* order as previously demonstrated, with **Special>Overlays>**Trajectory. The below also adjusts the scales of the Shepard plot by clicking on them (**X axis** & **Y axis** tabs). You might also enjoy repeating this 2-d MDS with, say, (Number of restarts: 5) & (Minimum stress: 0.001) and (✓Animate) switched on. Make sure that you make the above symbol and label changes, and especially the trajectory overlay, before starting the animation, with ▶. Even with the rather small number of points here (21) and the clear pattern, note how often the convergence does get trapped in a sub-optimal solution. This is also clear from the results window, mMDS1, from the initial run with 50 restarts, with only 20-30% of those converging to the (probably) lowest stress of 0.235.

v7

v7

Marginally higher stress is to be expected from an *m*MDS solution than an *n*MDS one, even if the Shepard diagram does show a linear fit through the origin to be an excellent description of the relationship of dissimilarities to plot distances – this is because *n*MDS's monotonic regression can make fine-scale steps to 'chase' the data even when displaying essentially a straight line (especially if there are few samples and therefore not too many points on the Shepard plot to 'chase'). But a high stress of 0.24 for *m*MDS, compared with a low one of 0.09 for *n*MDS, shows a drastically poorer fit for the former, and the reason is clear from the Shepard diagram: the points do form a fairly tight relationship with relatively low scatter (i.e. will fit well into a 2-d representation) but this is not linear through the origin, the assumption for metric MDS. In fact, the *m*MDS routine appears to be telling us that it can make good sense of these dissimilarities as a 2-d configuration, provided it ignores the model we have specified! It is the lack of fit to the model, rather than innate distortion in the 2-d solution (i.e. high variability in distance at each dissimilarity value), which is largely inflating the stress – this is also seen in the way the *m*MDS is similar to the earlier *n*MDS.

v7



Threshold metric MDS (*tm*MDS)

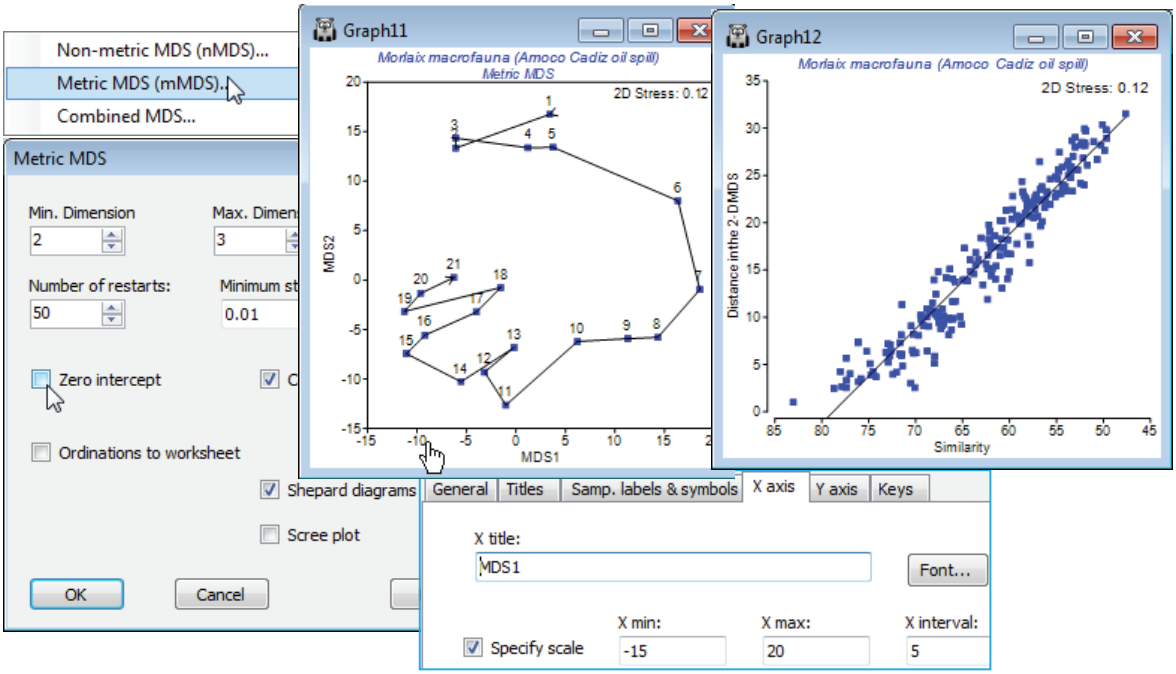
v7

The *m*MDS Shepard diagram above strongly suggests a middle course between simple metric and non-metric MDS, which we shall term *threshold metric MDS* (*tm*MDS), in which the Shepard plot is fitted by a straight line but not through the origin. It will almost always (as here) intersect the *x* axis of the Shepard plot, i.e. dissimilarities below that threshold are expected to be represented by zero distances – coincident points in the configuration. This intersection with the *x* axis is therefore an interesting parameter, essentially reflecting sampling error – the dissimilarity among replicate samples from the same condition (same site, time etc). For many data sets, metric MDS is unable adequately to represent both the genuine structure in the samples and the sampling variability amongst replicates in a low-d plot, without distorting linearity of the dissimilarity scale interpreted directly as distances. *n*MDS solves this by compressing the lower dissimilarities (often replicate error) into a much smaller range of near-zero distances, so that there is 'room' in 2-d to display the

v7

real structure among groups and along gradients etc. *tm*MDS solves it more crudely by attempting to truncate all dissimilarities below a fitted threshold to zero, subtracting the sampling error (which is represented by this threshold) from the larger dissimilarities. The advantage of *tm*MDS is that the ordinations will still have meaningful axis scales, in terms of the original dissimilarities – though it should be remembered that the threshold dissimilarity on the Shepard plot needs to be added back to any distances 'read off' from the ordination scales in order to produce a dissimilarity.

Run *tm*MDS on Resem1, by selecting **Metric MDS (mMDS)** and unchecking (✓Zero intercept). Refine the axis scales of the *tm*MDS plot from the **X axis** tab (reached from e.g. **Graph>General**).



v7

Note how variability around the line in the Shepard plot tightens up even further, now a reasonable model is fitted, and the *tm*MDS is virtually indistinguishable from the *n*MDS, seen much earlier in this section (and Fig. 5.8 in CiMC). The 2-d *n*MDS stress (see earlier Shepard plot or Fig. 5.9) is slightly lower at 0.09 rather than 0.12, a typical difference between a non-metric and metric stress in cases where both fit well (the relationship is genuinely linear). Given the low stress, the scales on the *tm*MDS can be used to give a 'feel' for the actual dissimilarity between different states of this community (rotating the plot may help to read off specific distances). The largest community difference between a sample prior to the spill and one in any subsequent year is about 30 units and the Shepard diagram shows the fitted threshold to be 21% dissimilarity (79% on the *x* axis scale of similarity). Adding these gives a dissimilarity of about 51%, and examination of the resemblance matrix shows the similarity between times 1 (April '77) and 11 (July '79) to be just under 52%. While *tm*MDS works well in this case it will not always be an appropriate ordination method, of course – any situation in which there is large scale turnover of species across the set of samples will tend to have a Shepard diagram which is strongly curvilinear at the large dissimilarity end of the scale, especially as it approaches 100%, and the ability of nMDS to model that relationship is crucial to an interpretable low-d display: small dissimilarities are compressed, large ones expanded.

Re-save the current workspace as Morlaix ws and close it.

Metric MDS for ordinating few points

Though usually greatly inferior to *n*MDS for typical ordinations from community similarities (e.g. with coefficients from the Bray-Curtis family), purely metric MDS (*m*MDS) becomes a valuable tool for ordinations based on only a handful of samples, when the information content of the ranks in a similarity matrix is often insufficient to constrain the problem. E.g. for 4 samples we have only the 6 numbers, 1, 2, 3, 4, 5, 6(!), and an infinity of stress zero solutions may be possible in *n*MDS. *v7* *m*MDS uses the actual (dis)similarities as the distances between points and solves a perfectly well-defined problem – even as few as 4 points in 2-d is unlikely to have an exactly zero stress solution. Such small-scale ordinations usually arise as means plots: averages are carried out for a number of groups over replicates or other crossed factors (e.g. averaging over times for a number of sites, or vice versa). This is achieved by **Tools>Average** on the transformed data matrix and calculation of the resemblances among these averaged samples (or, in the context of PERMANOVA+, by taking the menu item **Distances among centroids** on the original resemblance matrix, see Anderson et al, 2008, the PERMANOVA+ User manual). It is precisely such situations, with very few samples and thus few points on the Shepard diagram, where linearity of the distance *vs* dissimilarity relationship may be viable. Furthermore, the likelihood of the straight line going through the origin is increased, i.e. two means from exactly the same community structure will tend to have very low dissimilarity, because sampling variability is reduced by the averaging over replicates or other factors.

As an example, re-open the WA fish ws workspace of dietary assemblages of 7 species of Western Australian fish, for which averages over the (transformed) replicate gut samples from each species were calculated a few pages ago (in the presentation of segmented bubble plots). From the Bray-Curtis resemblance matrix computed on those averages, **Select>Samples>**(•Factor levels)>(Factor name: species)>**Levels** and Include only the four species with reasonably large numbers of samples – the three congeneric *Sillago* species (*S. schomb., S. bassen., S. vittata*) and *A. ogilbyi*, which all *v7* have between 10 and 16 replicate pools (each of 5 fish guts). Run **Metric MDS (mMDS)** on this 4 sample resemblance matrix. This only has 6 entries and *n*MDS would have insufficient information to be reliable (although not actually collapsing to a degenerate solution in this case, though it will often do so for 4-sample plots). *m*MDS, however, is seen to be valid – the 2-d Shepard diagram shows a reasonably convincing low-stress linear relationship, passing through the origin – and the 2-d ordination actually places these 4 points in very similar relationship to each other as seen in the previous *n*MDS plot for all 7 species (in that case, *n*MDS had plenty of information to work with, since there were 21 resemblances to rank). Note that the tidying up of this plot used: the **General** tab to uncheck (✓Plot history), increase (Overall font scale) and increase (Size) on **Info font** (stress value); **Samp. labels & symbols** to (Label✓By factor species full name) and to increase (Symbol> *v7* Size); both **X axis** and **Y axis** tabs to (✓Specify scale); and the **Keys** tab to uncheck (✓Plot key).

Resave the workspace as WA fish ws for use in the next section, and close it.

'Fix collapse' in nMDS

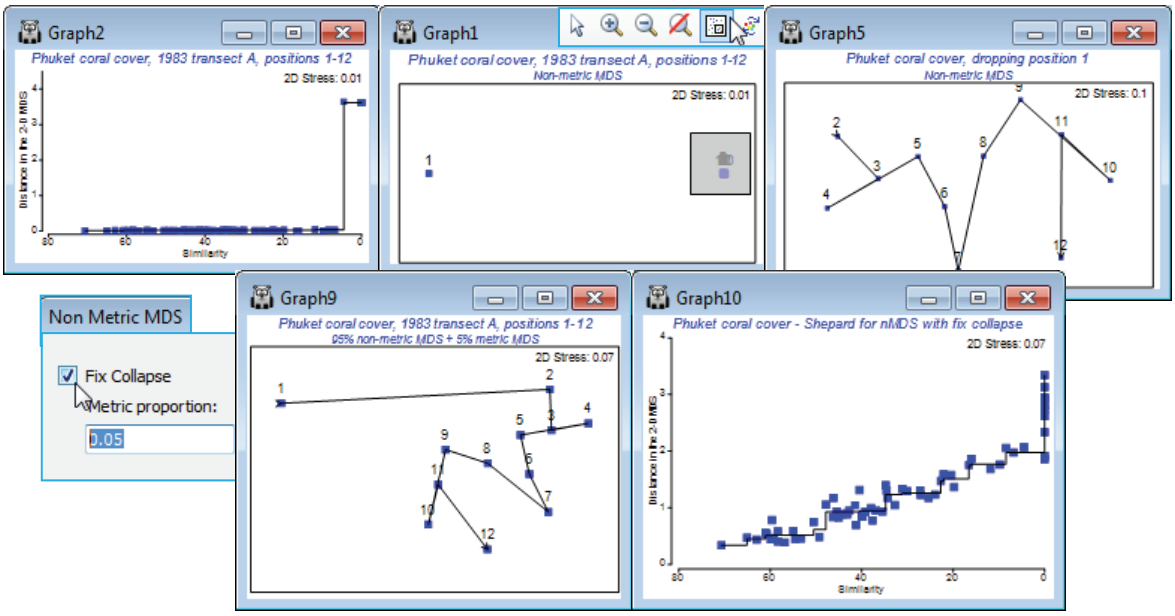
v7

The PRIMER 7 *n*MDS dialog box includes a new option to (✓Fix Collapse). This addresses the problem of indeterminacy in some *n*MDS plots, when two (or more) groups of samples have very different communities, such that all dissimilarities between the groups are larger than any of the dissimilarities within groups. There might then not be sufficient constraints in the rank information for *n*MDS to be able to locate these two groups in relation to each other. The two groups must be placed at least far enough apart to make their nearest neighbours further away from each other than the furthest neighbour distance within groups. But moving the groups further apart still, whilst keeping their internal layouts the same, may satisfy all the rank orders in the resemblance matrix in just the same way. In the successive steps of an *n*MDS iteration, the groups are gradually placed at ever increasing distance to each other, so the outcome is convergence to a two point solution where each group has collapsed to a point. The groups can be ordinated separately, having drawn a box around each, by **Graph>MDS Subset** (or sometimes just one box if, as below, the plot splits into a single outlying sample and the remaining points). But a better solution is to recognise that while *n*MDS has no measurement scale to determine the relative placing of the two groups, *m*MDS does (e.g. if the average dissimilarity between the two groups is 80% and the maximum within groups is 40%, the nearest neighbours between groups are placed at approximately twice the distance of the furthest neighbours within groups). It is not necessary to throw away all the flexibility of an *n*MDS, in capturing the within-group structures, by moving entirely to *m*MDS – the better solution is to use a mixture of the two stress functions. The default, if the (✓Fix Collapse) option is selected in *n*MDS, is to mix only 5% of metric stress with 95% non-metric stress: (Metric proportion: 0.05). The result does not appear to be at all sensitive to the choice of this proportion; all the *n*MDS needs is just enough information from the resemblance scale to fix the relative placing of the groups.

(Ko Phuket transects of coral reefs)

Live cover of a coral reef assemblage was recorded from 'plotless line-samples' (of 10m length) perpendicular to, and at 10m spacing along an onshore-offshore transect (A) at Cape Panwa, Ko Phuket, Thailand. (Samples taken in 1983-88 are described in Clarke KR, Warwick RM & Brown BE 1993, *Mar Ecol Prog Ser* 102). We shall meet these data more extensively later, but for now, open Phuket coral cover 83-87.pri in C:\Examples v7\Phuket corals, and select only the first year of sampling, 1983 (**Select>Samples**, as seen on the previous page, on the Year factor). Transform by square root and take Bray-Curtis resemblances of this 12-point transect (factor Position). An *n*MDS without the (✓Fix Collapse) option is seen to collapse to a single point (position 1, closest to the shore) and the remaining points – note in the Shepard diagram that a major step or steps like this always indicate a degenerate solution. An **MDS Subset** could be performed on positions 2-12, as shown below, but repeating all 12 points in an *n*MDS with the (✓Fix Collapse) option gives a more satisfactory (and low stress) Shepard diagram and a better description. The serial change in coral communities over the positions is best seen by **Graph>Special>Overlays>**(✓Overlay trajectory).

v7



Save the workspace as Phuket ws for use in Section 9, and close it.

Combined *n*MDS

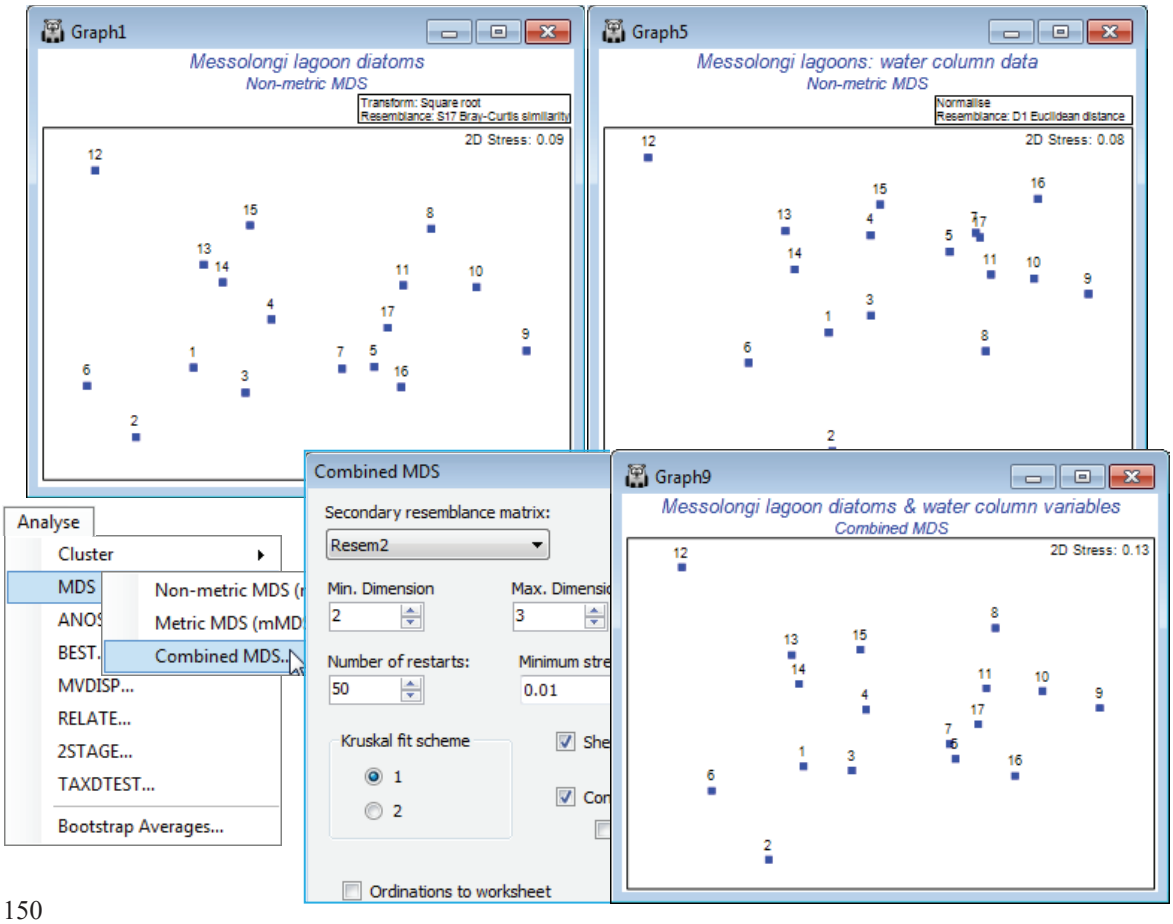
v7

A variation of the above, and a final new PRIMER 7 feature to discuss in this section, is the option to mix, in equal measure, stress functions from two resemblance matrices, in a single *n*MDS. There are possible applications to data sets in which a single MDS is required from two different types of data, which may need separate dissimilarity definitions, e.g. perhaps in a 'biotope' ordination, data might consist of biotic communities, usually analysed with a Bray-Curtis similarity, and continuous environmental variables which would typically be normalised and input to Euclidean distance. The practical reality is that it would still be better to somehow build those into a common matrix, which could be analysed with the same resemblance measure, because then all the analyses based on a single triangular matrix (clustering, ANOSIM tests etc) would become available. But if the sole purpose is to produce an ordination which combines two resemblance matrices for the same set of sample labels (same sites, times etc) then **Analyse>MDS>Combined MDS** offers this possibility. The dialog is more or less the same as that for *n*MDS except for specification of the (Secondary resemblance matrix:) to mix with that for the active sheet. A solution is sought, in the usual range of possible dimensions, by minimising an equal mix of the two stress functions, and the combined MDS is displayed along with the Shepard diagram for resemblances from the active sheet (to get the other Shepard diagram, re-run the routine with the 'secondary' resemblance matrix now as the active sheet – the stress value quoted on both plots will be the overall average stress). It is inevitable that the overall stress will be higher than that for the two analyses separately, because the routine is trying to compromise between a configuration of samples which would best suit each matrix separately; the stress only matches those of the individual solutions when they are identical.

(Messolongi diatoms & abiotic data)

A study of water column diatom assemblages (193 species) at 17 sites in the Messolongi lagoons of E Central Greece, and a matching set of 11 water quality variables measured at the same sites (with salinity fluctuating between brackish water and fully marine, and nutrient enrichment within certain lagoons) is looked at in detail in Section 13, on methods for biota-environment linkage. If analyses given there for the biotic data (root transform of diatom densities, followed by Bray-Curtis) and abiotic variables (log transform for all water column chemistry variables, excepting Temperature, Salinity, DO_2 and pH, followed by normalisation and Euclidean distance) are repeated here, they lead to the separate *n*MDS plots shown below, with respective stress 0.09 and 0.08. There is a remarkable similarity in the patterns of the 17 sites between the two analyses, and running **Analyse >MDS>Combined MDS** on the Bray-Curtis similarity, supplying the Euclidean distance as the secondary resemblance matrix, leads to the combined *n*MDS shown, with stress of 0.13.

v7



9. Analysis of Similarity tests (unordered and ordered *ANOSIM*)

ANOSIM introduction

v7

The series of ANOSIM (analysis of similarity) tests, accessed through **Analyse>ANOSIM**, operate on a resemblance matrix as the active sheet and carry out non-parametric tests for designs which broadly parallel univariate 1- and 2- way ANOVA (analysis of variance) crossed and nested cases, extended in PRIMER 7 to cover all combinations of crossed and nested 3-way designs. At the simplest level, a one-way layout of groups (e.g. of different times or sites or treatments), with replicates within each group, allows a test of the null hypothesis that there are no assemblage differences between the groups against an alternative which specifies that there are – but makes no assumption about the nature of those differences. However, another new feature in PRIMER 7 is the addition of a test of the same null hypothesis but against an alternative which specifies that the groups differ in a predetermined order, for example exhibiting a serial time trend or a continuous community change on approaching a point source impact. These ordered ANOSIM tests generalise, in a very natural way, the standard one-way ANOSIM R statistic to R^O (superscript O for Ordered), defined in Chapter 6 of CiMC, and such ordering can be specified for all or any combination of factors in the 2- and 3-way designs. By narrowing the scope of the alternative hypothesis that the null is tested against, a greater degree of sensitivity (power) is obtained – though at the price of little or no sensitivity to detect group differences which do not conform to the hypothesised serial pattern. A crucial point to make is that the group designations (and specification of the group order under the alternative hypothesis) are made *prior* to seeing the data. ANOSIM is not a valid test of differences between groups generated by a cluster analysis, or other inspection of the data, otherwise the argument becomes entirely circular – use SIMPROF for these latter situations.

v7

v7

The simple 1-way R statistic is readily extended to a 2-way crossed layout, for example testing the null hypothesis that there are no differences between treatments (factor A), allowing for the fact that there may be site differences (factor B), in a case where all treatments are replicated at each site. Two-way crossed designs (A×B) are symmetric, so the procedure can be reversed to give also a test for the hypothesis of no site differences given that there may be treatment differences – the routine gives both sets of tests automatically. The other 2-way option is a 2-way nested design, where the two factors are hierarchical, B(A), for example a top-level factor of treatment differences A (control *vs* impacted areas), with a second factor of different sites B within each treatment, and representative replicate samples from each site. A test can be carried out for significant differences between sites within a treatment, but at the next hierarchical level up, the primary interest would be in testing for assemblage differences due to treatment. This compares treatment differences against assemblage variation among sites within a treatment, rather than among the sample replicates at a site, and there is an example of this in Chapter 6, CiMC. A different style of test is required in the case of a 2- (or 3-)way crossed layout when there are no replicates (or no genuine replicates and it is wise to average the *pseudo-replicates* for each of the factor combinations). PRIMER 7 will automatically recognise situations in which replicates are not available and attempt to calculate a different test statistic ρ (in whatever combinations of 2- and 3-factor designs it is encountered).

v7

v7

Three-way crossed and nested designs, also now covered in PRIMER 7, are of the types: A×B×C, fully crossed; C(B(A)), fully nested, C within B within A; C(A×B), C nested in all combinations of A crossed with B; and B×C(A), B crossed with C, where C is nested in A. These routines are all permutation tests, making a bare minimum of assumptions and consistent with the philosophy of the PRIMER routines that the primary information on relationships among samples is summarised in the ranks of the resemblance matrix – the basis for the preferred ordination technique of non-metric MDS. The tests apply to any resemblance matrix, so are equally effective at testing for assemblage change on biotic similarities, environmental change on Euclidean or other distances, change in biomarkers, particle sizes etc. None of the construction or concepts underpinning these tests is covered in this Section – that is all in the extensive Chapter 6 of CiMC, which includes the detailed Tables 6.3 and 6.4, listing the precise test statistics (and whether pairwise tests are possible when there are more than two levels of a factor) for every combination of: 1-, 2- and 3-way tests; unordered or ordered alternative hypotheses; and with or without replication. (That chapter finishes with some comments on the limits of construction of purely non-parametric tests, e.g. the non-existence of tests for interaction – a metric concept and the springboard for PERMANOVA). The below simply shows examples of the different tests and how the results windows are interpreted.

v7

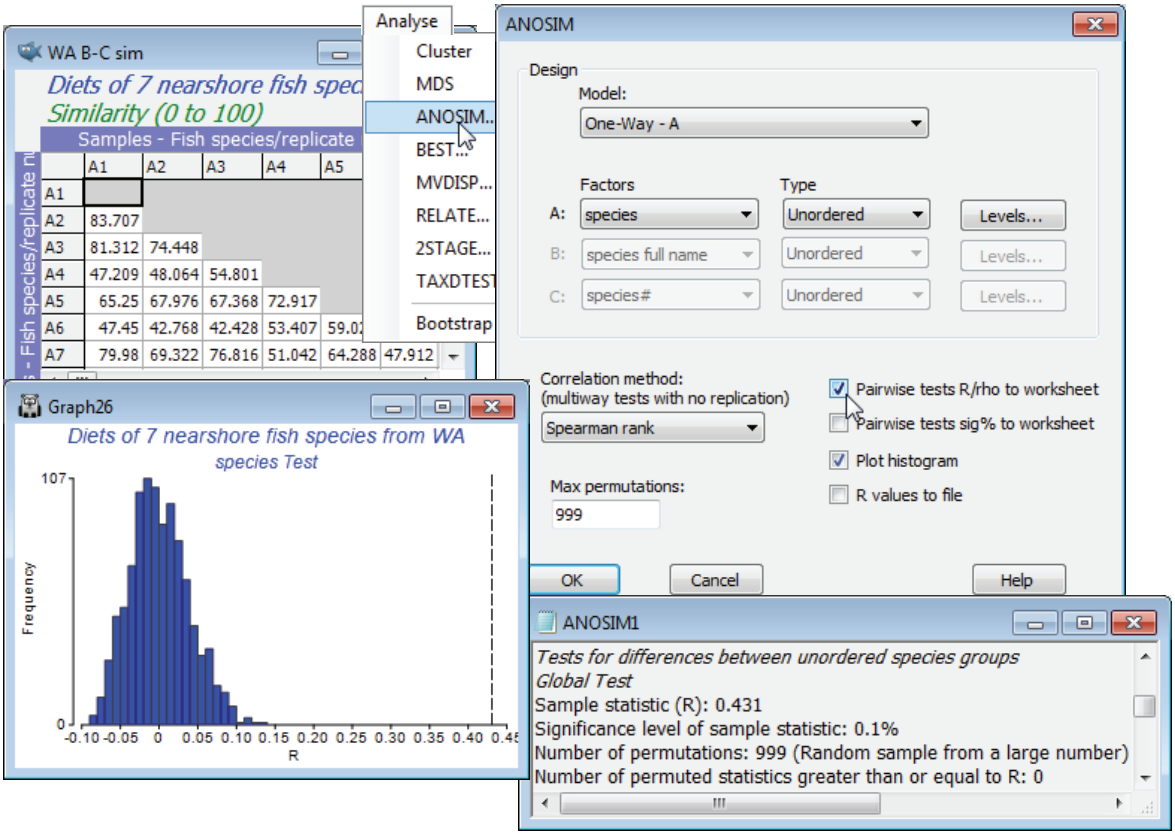
1-way layout (WA fish diet example)

Return to the W Australian fish diet data, introduced at the start of Section 4 and last seen under the Higher-d & scree plots heading of Section 8, showing MDS plots of dietary categories found in 65 (pooled) gut samples from 7 fish species. If the workspace WA fish ws is not available, re-open the data WA fish diets %vol.pri in C:\Examples v7\WA fish diets, exclude the samples A9, B3 and B4 (justifiable on the grounds of very low total gut content), standardise the samples, transform them with square root and take Bray-Curtis similarities, renaming this WA B-C sim. Re-run the *n*MDS and note that the number of replicates – the pools of fish from each species – is very uneven (from 3 to 15), as is the variability in diet for the different species (the dispersion of samples in the MDS space). Assumptions of balanced replication and the equivalent of ANOVA variance homogeneity are clearly not met here, but the ANOSIM test does not require such assumptions for its validity. Approximate balance in replication is still a good idea because it enhances the sensitivity of the tests, and comparable multivariate dispersion within each group makes interpretation simpler, but neither is possible here. ANOSIM tests the hypothesis that there are no dietary differences of any sort among the fish species. This null hypothesis can be rejected either because species require different food sources or because some have a much more variable diets than others, though they may feed on some of the same items – either or both reasons may contribute to rejection of the null.

From the active window WA B-C sim, take **Analyse>ANOSIM>**(Model•One-Way - A)>(Factors A: species)>(Type Unordered), since clearly there is no prior expectation that if the diets differ then they can only do so along a steady gradient of change in a particular order of the fish species. (This is, however, exactly the expectation we have later when looking at the dietary changes of a single fish species at different stages of maturity, when the logical test will be an Ordered one). On the other choices, take (✓Pairwise tests R to worksheet) but leave the default settings for the remaining options, e.g. (Max permutations: 999) and **Levels** specifies that all 7 fish species are to be included. Three windows are created in the Explorer tree. *ANOSIM1* is the results window, specifying the sheet and factor on which the test is performed, and giving the results of the overall ANOSIM test of the hypothesis of no differences in diet among any of the fish species, followed (when there are more than two groups, as here) by pairwise tests between the diets of every pair of fish species. The second output is a histogram of the permutation distribution of the ANOSIM test statistic, R, under the null hypothesis of the global test (though note that it is not the correct permutation distribution for any of the pairwise tests). The third output, requested here, is the set of observed pairwise R values held as a triangular (resemblance) matrix, and a similar option exists (not taken) to place the % significance levels for this set of pairwise tests in a further resemblance matrix.

v7

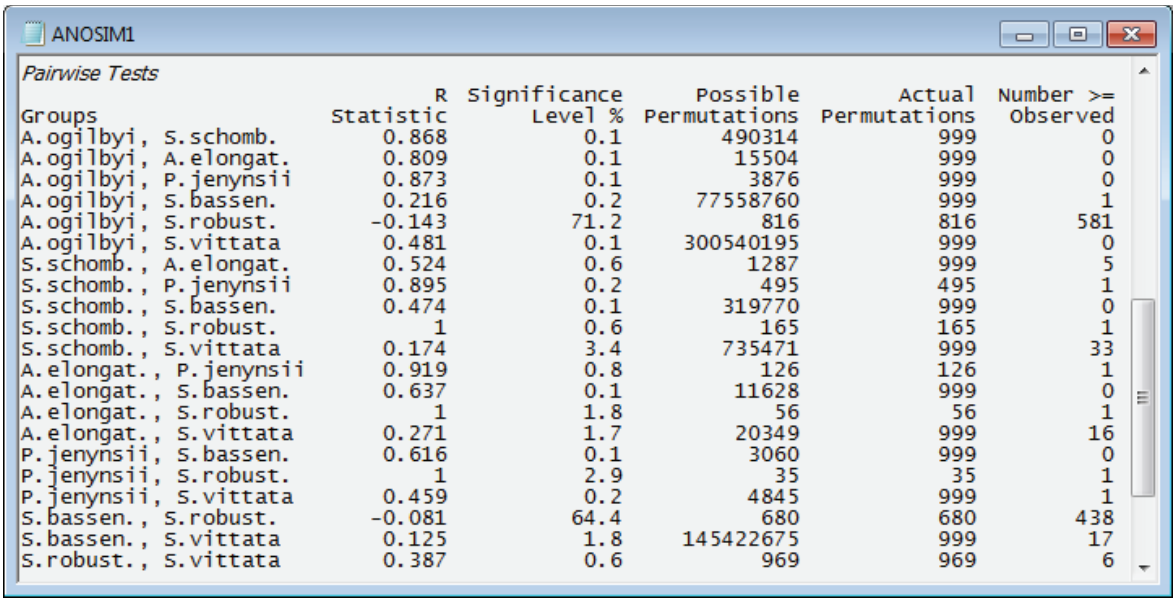
v7



This histogram is centred around zero – if there are no dietary differences then the average rank resemblance among and within groups will be much the same, and R (based on the difference between these two averages, see Chapter 6, CiMC) will be near zero. It can rise a little above (or below) zero by chance, when there are no differences among diets, but the histogram shows that it will never get larger than about R = 0.15. The true value of R for these data is also shown, as a dotted line, namely R = 0.43, and this is clearly much larger than any of the 999 permuted values, causing rejection of the null hypothesis at a significance level of at least 1 in 1000 (P<0.001, or as the PRIMER output prefers it: p<0.1%). The same information is repeated in the results window ANOSIM1 under the heading *Global Test*, namely the overall observed R statistic of 0.431, its significance level (p<0.1%), how many permutations were computed in order to determine this (999), and how many of those permutations gave an R value as large, or larger, than the observed R of 0.431 (none). The total number of possible permutations – distinct ways of dividing the 65 samples amongst the 7 fish species, keeping the same number of replicates for each species – is extremely large and is therefore not displayed. In other cases, with few replicates, this third row will give the exact number of possible permutations, and if this is less than the specified (Max permutations:) in the ANOSIM dialog box, then R will be evaluated for all possible permutations. Setting (Max permutations: 9999) will increase the significance level here to p<0.01% (it is clear from the histogram that, almost irrespective of how many permutations are chosen, a value as large as R = 0.43 will not be obtained by chance, so the significance level for the global test of no dietary differences can be made arbitrarily small, by increasing the number of permutations).

Pairwise comparisons

The table ending the results window gives the pairwise comparisons. For each pair of groups (fish species), the first data column is of pairwise R statistics. These are again a difference of average rank dissimilarities between and within the two groups, scaled so that R varies between roughly 0: there are no differences, and 1: all dissimilarities between gut contents of different fish species are larger than any dissimilarity among samples within either species. The second column gives the statistical significance for a test of R = 0 (again as a percentage, so that p<0.1% means less than a 1 in 1000 chance). The number of possible permutations follows, then the number actually computed – 999 in most cases because the possible number is usually much larger than this, here. The final column gives the number of R values from the permutations that exceed (or equal) the real R in the first column, from which the significance in column 2 is calculated. [Note that there needs to be a slight difference in this computation depending on whether all possible permutations are evaluated. Thus row 1: *A. ogilbyi* v *S. schomb.*, R = 0.868, p<100(1+0)/(1+999) = 0.1%, whereas row 12: *A. elongat.* v *P. jenynsii*, R = 0.919, p = 100(1/126) = 0.8%. The second is clearest: the observed value of 0.919 is the most extreme of 126 permutations and thus has probability 1 in 126 of occurring by chance. In the first case, we do not observe the real value of 0.868 in our randomly chosen set of 999 permutations, but that does not make the probability p = 0/999 = 0. We know there exists one permutation which would give R at least 0.868 – the real configuration – and we have looked at 1000 permutations overall (the 999 random plus the real one) so the probability is < 1 in 1000.]


ANOSIM1

Pairwise Tests

Groups	R Statistic	Significance Level %	Possible Permutations	Actual Permutations	Number >= Observed
A.ogilbyi, S.schomb.	0.868	0.1	490314	999	0
A.ogilbyi, A.elongat.	0.809	0.1	15504	999	0
A.ogilbyi, P.jenynsii	0.873	0.1	3876	999	0
A.ogilbyi, S.bassen.	0.216	0.2	77558760	999	1
A.ogilbyi, S.robust.	-0.143	71.2	816	816	581
A.ogilbyi, S.vittata	0.481	0.1	300540195	999	0
S.schomb., A.elongat.	0.524	0.6	1287	999	5
S.schomb., P.jenynsii	0.895	0.2	495	495	1
S.schomb., S.bassen.	0.474	0.1	319770	999	0
S.schomb., S.robust.	1	0.6	165	165	1
S.schomb., S.vittata	0.174	3.4	735471	999	33
A.elongat., P.jenynsii	0.919	0.8	126	126	1
A.elongat., S.bassen.	0.637	0.1	11628	999	0
A.elongat., S.robust.	1	1.8	56	56	1
A.elongat., S.vittata	0.271	1.7	20349	999	16
P.jenynsii, S.bassen.	0.616	0.1	3060	999	0
P.jenynsii, S.robust.	1	2.9	35	35	1
P.jenynsii, S.vittata	0.459	0.2	4845	999	1
S.bassen., S.robust.	-0.081	64.4	680	680	438
S.bassen., S.vittata	0.125	1.8	145422675	999	17
S.robust., S.vittata	0.387	0.6	969	969	6

Interpreting these pairwise tables must be done with care. The significance level is very dependent on the number of replicates in the comparison. For example, row 4: *A. ogilbyi* v *S. bassen.*, p<0.2% (your value may differ slightly because each time the routine is run, different random permutations will be generated). This appears highly significant, but the R value is negligibly small, at 0.216. The test tells us that these two species probably do not have exactly the same diet (the hypothesis R = 0 can be rejected) but the R value tells us that the diets are strongly overlapping and barely differ (R is close to zero). This can happen, just as in ordinary univariate statistics, because the number of replicates is large for the two groups, giving 77 million possible permutations – biologically trivial differences can still be statistically significant when the test's power is large. In total contrast, row 17: *P. jenynsii* v *S. robust.*, p<2.9%, still significant but only just (at the 5% level), has an observed R of 1.0, the largest possible value, which shows completely different diets. Such a large value of R does not give a small value of p because there are only 35 possible permutations (few replicates in both groups). Which is therefore the most useful column to interpret? It has to be the R values and not the p values. R is largely not a function of the number of replicates (i.e. possible permutations) but an absolute measure of differences between two or more groups in the high-dimensional space of the data, whereas p is always hijacked by the sample size. It is for this reason that PRIMER does not implement a Bonferroni-type correction on its pairwise significance levels – it gives an illusion of certitude which is not justified. The global test of any differences between groups is important: if the null hypothesis is not rejected then the user has no licence to look at the pairwise comparisons. However, if the global test strongly suggests that there are differences worth examining, the focus shifts to the pairwise R values – large values there indicate where the major differences are found.

Other 1-way ANOSIM options

Checking the (✓Pairwise tests to worksheet) box has also sent the above R values to a worksheet in triangular format, which could be a useful layout for tabulating ANOSIM results in a publication. More subtly, this can be regarded as a resemblance matrix (of distance-type) in its own right – the higher the value of R the greater the separation of replicates from two groups in the high-d (prey) space. Inputting this to an MDS plot will display the relationships between these 7 groups, and can be seen as a type of means plot. [Note that this triangular array is not a sensible distance matrix at present because it can, and does, contain (small) negative values. Input to metric MDS without some prior rescaling would be problematic therefore. However, *n*MDS effectively works only on the rank orders of the entries so there is no need to rescale them – the lowest values (the negative ones) indicate the least established differences in diet and the highest values (R=1) the greatest differences, which is exactly what is required for a sensible *n*MDS plot here. Dropping the negative signs, by taking absolute values of the entries, would not be the technically correct approach here.]

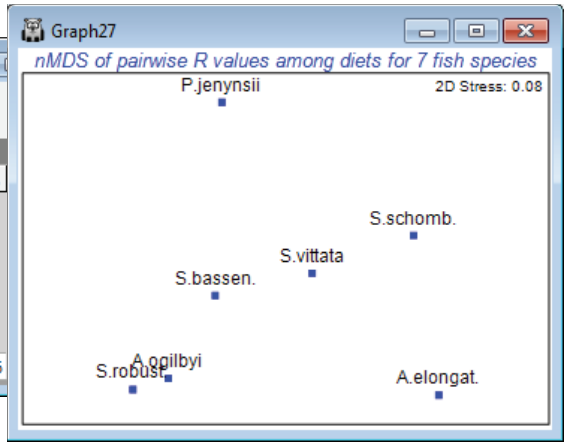
v7

The more straightforward means plot is, as we have seen before, is to average the replicates, and then calculate Bray-Curtis between these mean dietary samples, ordinating by *n*MDS or *m*MDS. But there are many other possibilities for a direct means plot! The data could be transformed before or after averaging, or the dissimilarities could be averaged – or even their ranks averaged. PRIMER 7 now has the option to average (dis)similarities across a group structure, with **Tools>Average** for an active window of a resemblance matrix. **Tools>Rank distance** will also replace resemblance entries with their ranks. (A further option is given in the PERMANOVA+ add-on, of computing *distances among centroids* in the high-d PCO space formed from the resemblance matrix). These will all give means plots with slightly different emphases. In the case of the matrix of R values, this highlights relative group separations, i.e. adjusting differences by within-group dispersion.

Resem4

species pairwise tests R
R (-1 to 1)

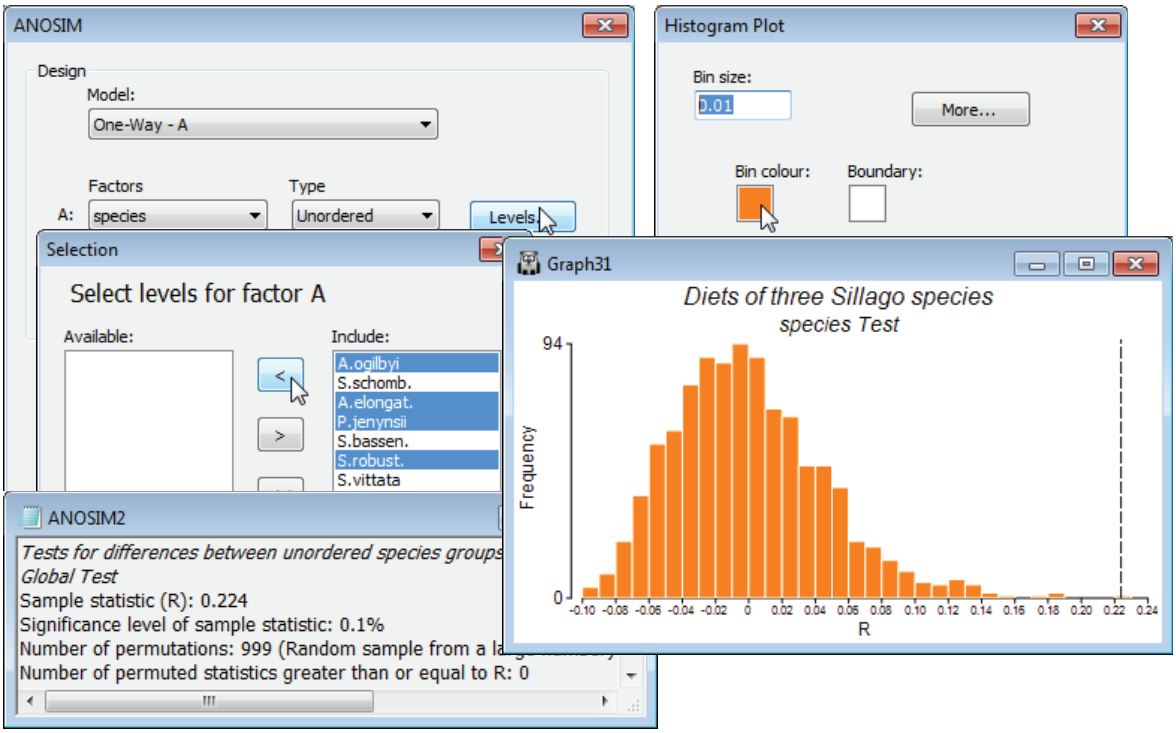
Other	A.ogilbyi	S.schomb.	A.elongat.	P.jenynsii	S.bassen.	S.robust.
A.ogilbyi						
S.schomb.	0.86805					
A.elongat.	0.80893	0.52368				
P.jenynsii	0.87327	0.89522	0.91875			
S.bassen.	0.2164	0.47366	0.63706	0.61635		
S.robust.	-0.1428	1	1	1	-0.081054	
S.vittata	0.48059	0.1742	0.27115	0.45933	0.12542	0.38686



v7

Other options within the ANOSIM routine include the ability to manipulate the histogram for the global R statistic by rescaling axes, titles etc (the usual **Graph>Sample Labels & Symbols** menu) and changing bin widths and, in v7, bin colours (**Graph>Special**), as for any other histogram plot. There is also a check box in the ANOSIM dialog to send (✓R values to file). You would then need to supply an *.txt file name which will hold a simple list – one number to a line, in simple text – of the R values for the 999 (or however many) permutations carried out for the global test. This would allow the null distribution data to be replotted, for example, in another statistical/graphical package.

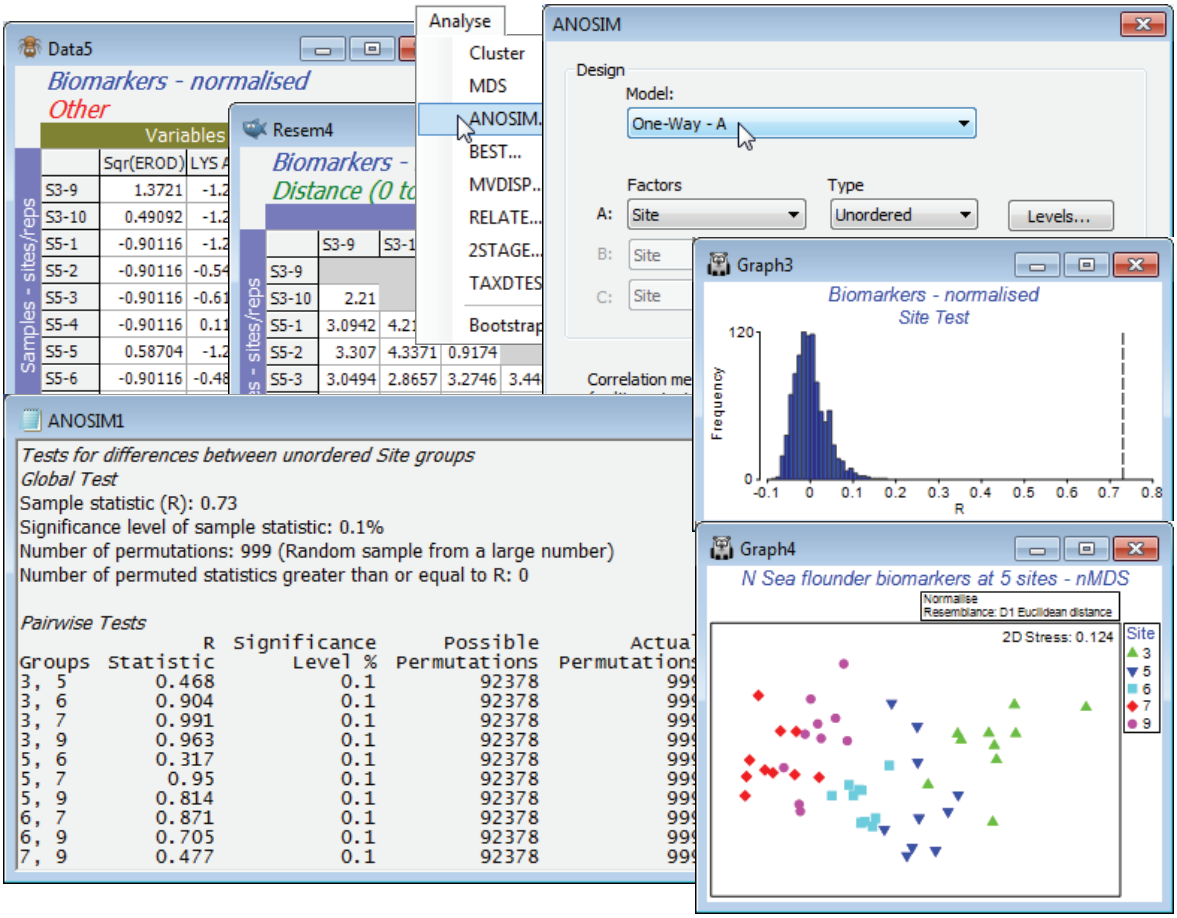
As noted earlier, the plotted histogram (and the listed R values) refer only to the global test for no differences among any of the groups. If you require a histogram for a specific pairwise comparison then you will need to pick out that pair of groups and re-run **ANOSIM**, selecting either externally, by **Select>Samples** on the original resemblance matrix, or internally, using the **Levels** button for A on the ANOSIM dialog. Both lead to the usual Selection dialog. For a pairwise test, it will make no difference to the R value (or to its significance level) whether the results are read from the above pairwise table or recalculated with just those groups selected, so this would only be useful: a) if you required the pairwise histogram, or b) a test for a specific subset of three groups, four groups etc was needed. As seen in Section 3, a relevant *a priori* hypothesis here concerns whether there are detectable dietary differences between the three congeneric *Sillago* fish species (*S. schomburgkii, S. bassensis* and *S. vittata*). After testing this, save and close the workspace WA fish ws.



1-way layout (Biomarkers example)

ANOSIM applies equally well to data on environmental, biomarker or morphometric variables, which might be transformable to approximate normality; it is then a robust alternative to classical multivariate (MANOVA) tests such as Wilks' lambda. The inevitable slight loss in power of the non-parametric test, if the data really were multivariate normal (and had few enough variables in relation to sample sizes to allow proper estimation) is more than compensated for by its robustness, general applicability, and lack of assumptions such as constant variance-covariance structures.

Re-open the N Sea ws workspace seen in Sections 4 & 5, with datasheet N Sea flounder biomarkers of 10 replicates from each of 5 N Sea sites, S3, S5, S6, S7, S9 (directory C:\Examples v7\N Sea biomarkers). Of the suite of 11 biomarkers it was previously suggested that EROD and LIPID VAC might benefit from square root transformation, being modestly right-skewed – earlier performed by highlighting them and **Pre-treatment>Transformation(individual)**>(Expression: SQR(V)). The variables are put on a common measurement scale with **Pre-treatment>Normalise Variables** and an appropriate resemblance calculation is **Analyse>Resemblance>**(Measure•Euclidean distance), before **Analyse>ANOSIM** on factor Site, as above. The results show a significant (mainly large) separation of the biomarker responses at all sites, seen clearly also in an *n*MDS plot (or *m*MDS would have nearly as good a stress here). Resave and close the workspace, N Sea ws.



In this example, the sites were along a transect from the mouth of the Elbe (S3) to the Dogger Bank (S9) but there was no strong presumption in advance of the data collection that responses of the biomarker suite, if present at all, must conform to a monotonic gradient of change. This is not a clear case of a decreasing contaminant gradient with distance from a point source impact, since over such a large geographic region (and with motile organisms) many environmental drivers may be in play, only some of them anthropogenic. An ordered ANOSIM test would then run the risk of failing to detect differences among sites if these are not in the physical order of the sites along the transect. And, whilst the above *n*MDS does then show responses that are mainly aligned with the transect – and the ordered R^O gives a similar and also highly significant value of 0.74 compared to the unordered R = 0.73 statistic – this pattern is not entirely consistent (e.g. sites 6, 7 and 9). Of course, a decision about which test to use should be made prior to seeing any data, and it would have been unwise to opt for the ordered test and rule out any possibility of detecting community changes in which, for example, both ends of the transect were impacted in the same way (and therefore similar to each other) but the mid-transect sites reflected a different background state.

v7

1-way ordered ANOSIM (Ekofisk oil-field study)

However, for the Ekofisk oil-field study of the last section there is a postulated cause for benthic community change (the presence of an oilfield) and impacts, if any, are expected to result in a monotonic change with increasing distance from the drilling centre. For the four distance groups, defined prior to the data analysis: D (<250m), C (250-1000m), B (1 - 3.5 km) and A (>3.5km), a test of the null hypothesis H_0:A=B=C=D against the ordered alternative H_1:A→B→C→D (or vice-versa) is entirely appropriate, and inability to detect the above scenario (sites near and far from the rig being similar but intermediate sites differing) would usually be a price well worth paying in pursuit of a more powerful test of this specific (ordered) alternative hypothesis. [We return to this point, and example, in Section 14, based on the *seriation with replication* test of Somerfield PJ, Clarke KR & Olsgard F 2002, *J Anim Ecol* 71:581-593, which tackles the same problem with a slightly different (*RELATE*) statistic, ρ. The relationship of the new, ordered ANOSIM statistic to that for the previous **Analyse>RELATE** test is covered in Chapter 6 of CiMC].

v7

v7

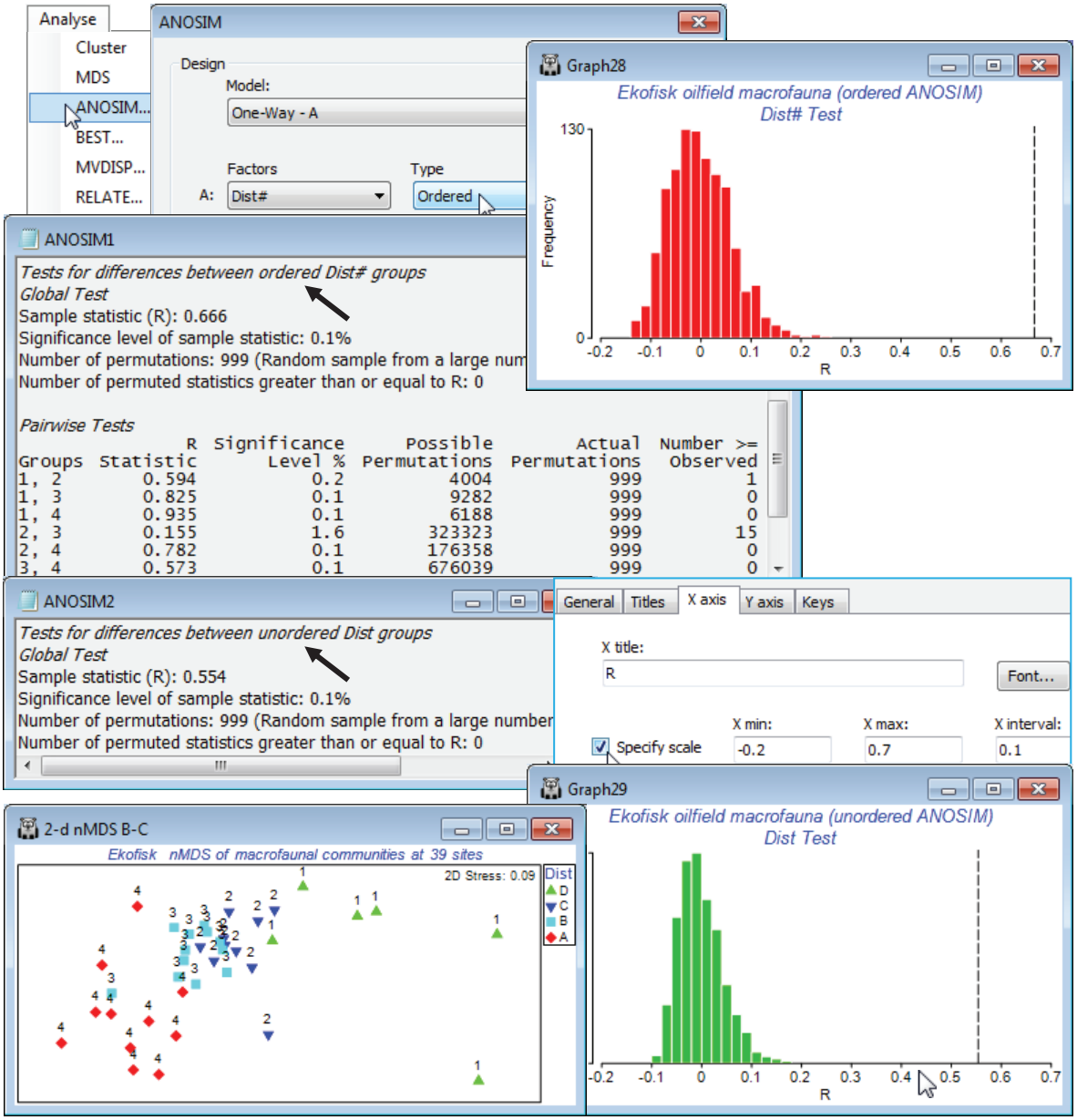
Re-open the Ekofisk ws workspace of Section 8, in which the data Ekofisk macrofauna counts from C:\Examples v7\Ekofisk macrofauna has been square-rooted and input to Bray-Curtis calculation,

v7 giving resemblance matrix B-C on sq rt. On this, take **Analyse>ANOSIM>**(Model•One-Way-A)> (Factors A: Dist#)>(Type Ordered). Note that the numeric form Dist# (levels 1-4) of the factor representing distances from the oilfield is required for the Ordered ANOSIM test; the alphabetic form Dist (levels D-A) is not recognised by PRIMER as an ordering (if Dist# is not available, re-

v7 create it by **Edit>Factors>Add>**(Add factor named: Dist#), and enter 1 at the top, opposite the first D, 2 opposite the first C, 3 when it changes to B, and 4 to A, then highlight the column and **Fill>Value** to fill in all the blanks appropriately).

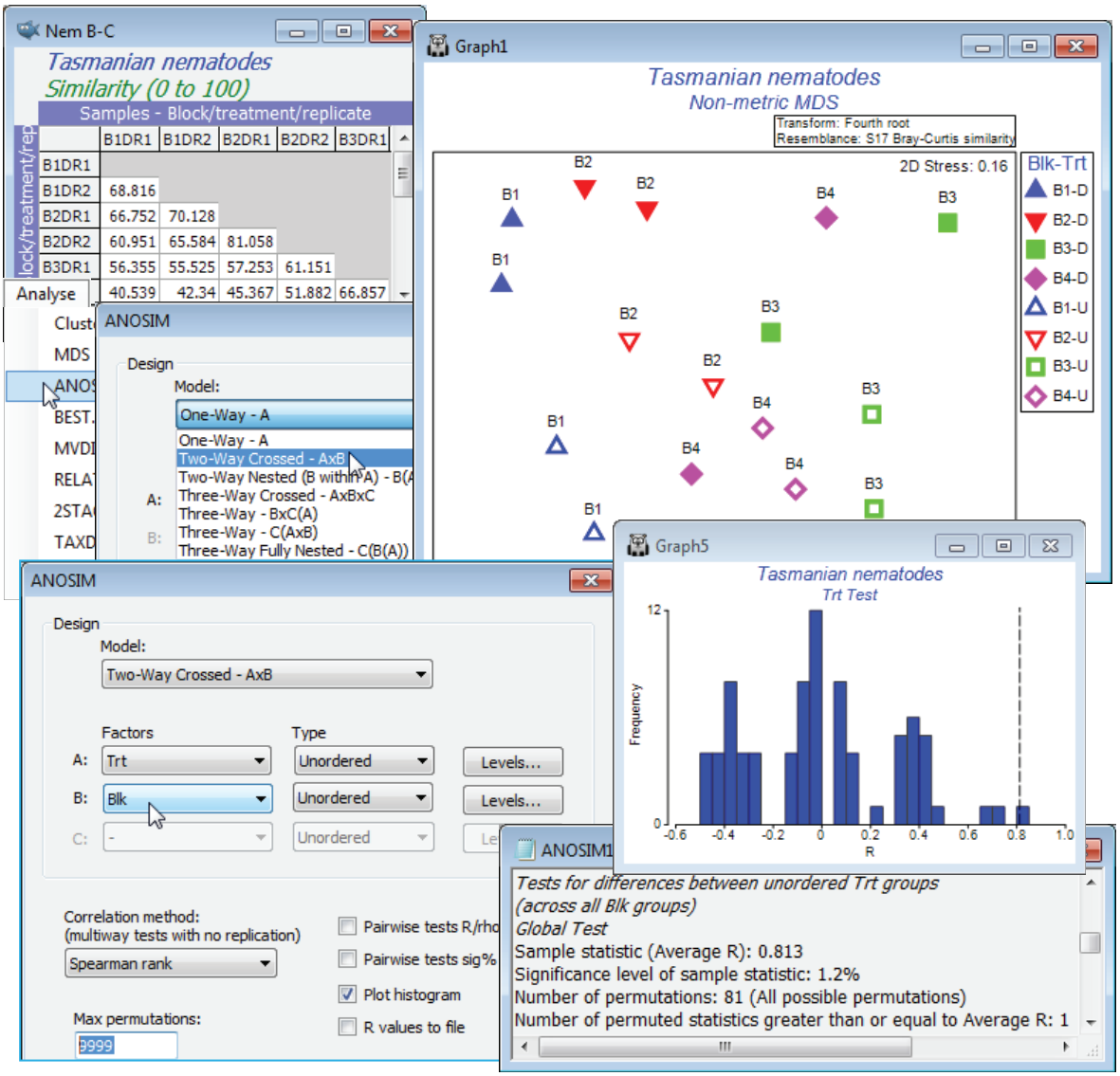
The outcome is an ordered R^O of 0.67 which, as can be seen from the null distribution histogram, unquestionably rejects the null hypothesis ($R^O = 0$) on a p<0.1% level test (i.e. P<0.001), and would do so for massively smaller values of p. The ensuing pairwise tests show a clear pattern of increasing R for increasing separation of the distance groups, exactly as one would expect from a serial community change. Only the differences between groups 2 and 3 (C and B) are at all borderline, with a low R of 0.16 (but still significant in conventional terms, at the 2% level). Note that the pairwise tests are always ordinary R statistics – there is no difference between an ordered and an unordered test when there are only two groups! Now re-run the ANOSIM specifying the unordered case, this time using either the alphabetic or numeric factors Dist or Dist# (it no longer matters). The R = 0.55 value, though still massively significant, is seen to be lower than the $R^O = 0.67$ of the

v7 ordered test. These statistics are directly comparable, and show the better fit of the underlying model (see Chapter 6, CiMC) of equi-stepped serial change than that of equi-different groups. This is clearly evident also from the previous *n*MDS plot. Save and close the workspace Ekofisk ws.



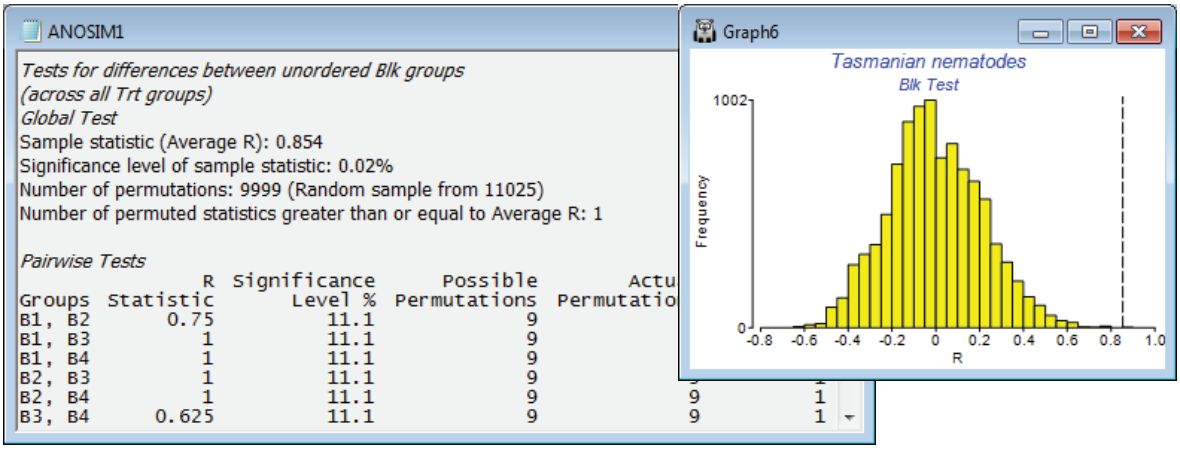
2-way crossed ANOSIM (Tasmanian crabs study)

An example of a 2-way crossed layout was introduced in Section 1, for meiofaunal communities in sediment patches either disturbed or undisturbed by soldier crabs (treatment factor Trt, levels D/U, factor A), over four areas of Tasmanian sandflat (block factor Blk, levels 1-4, factor B), with two replicates for each of the 8 combinations. Setting up of these factors was described in Section 2. Open (or create) the workspace Tasmania ws with datasheet Tasmania nematodes, take a fourth-root transform and compute Bray-Curtis similarity, renamed Nem B-C. Run this through *n*MDS and note the way the samples split rather convincingly between the effects of the different regions of sandflat (blocks, roughly across the page) and disturbed or undisturbed (treatments, roughly up the page; click on the Blk-Trt key to change symbol type/colour for blocks and open/closed for treatments). There are so few replicates, however, that this is not clear-cut and does need testing. Also, stress is quite high (about 0.16) so the picture may be misleading, and the test needs to be in the full-dimensional space represented by the resemblance matrix, as ANOSIM tests always are. Two-way crossed ANOSIM is carried out for the null hypotheses, H_0: no treatment effect, allowing for the fact that there may be differences between blocks, and also symmetrically for H_0: no block effect, allowing for the fact that there may be treatment effects. The test statistic for treatments is now the average of the 1-way ANOSIM R values for testing the treatments separately within each block, and the permutation procedure is also a constrained one, within blocks (see Chapter 6 and Fig. 6.7 of CiMC, which analyses the same study but for the full meiofaunal data – nematodes and copepods combined – for which the outcome is even clearer). Here, with Nem B-C as the active sheet, take **Analyse>ANOSIM>**(Model: Two-Way Crossed - AxB)>(Factors A: Trt & B: Blk) and both factors are treated as unordered (the treatment only has two levels anyway, and though the block areas will vary in their nematode assemblages there is no expectation that this will be on a strong gradient). Again, note that we could have restricted the analysis to use only some of the levels of either factor, with the **Levels** buttons, though this is not appropriate here.



For the test of the treatment effect, the histogram shows a typical null distribution for the ANOSIM test statistic when there are few replicates. With only 81 distinct permutations permitted (these correspond to all ways of simultaneously exchanging the four replicates within each block), the range of values that the statistic can take when there is no treatment effect is not at all smooth. This demonstrates why the null permutation distribution has to be recreated for each new data set and cannot rely on standard tables or distributional forms. And clearly there is no point for this test in increasing to (Max permutations: 9999), as we have done here – there are only 81 permutations and ANOSIM does them all. Nonetheless, the results in ANOSIM1 show that the observed statistic for testing treatments (global R = 0.813) is the largest obtainable for the 81 permutations, so gives a significance level of 1 in 81 (p = 1.2%). This would normally be considered sufficient to cast doubt on the null hypothesis of no treatment effect. If there were more than two treatments, the global test would be followed by pairwise comparison of treatments, exactly as for the 1-way ANOSIM case.

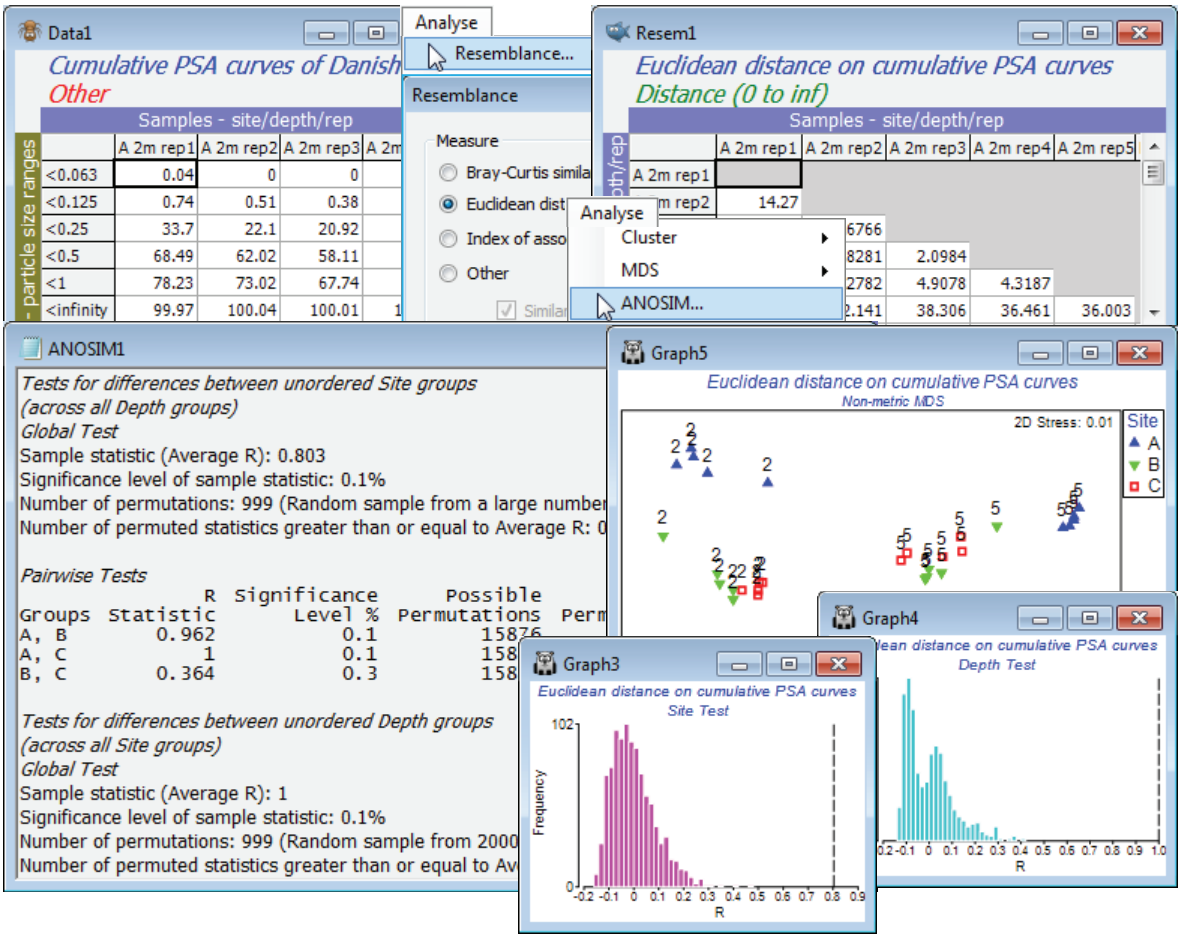
The second plot is much smoother because there are 11025 possible permutations of the replicates across blocks within each treatment, corresponding to the null hypothesis of no block effect, and a random subset (with replacement) of most of them has been evaluated. In fact, having seen that there are only 11025 to find, it would make sense to repeat the analysis, setting (Max permutations: 12000) – or any number >11025 – because ANOSIM will then compute the full set of 11025. But this was not done here, and the observed average R of 0.854 was seen to be the most extreme of all but one of the 9999 permutations examined – almost certainly that one was the real configuration but we cannot guarantee that if not all the permutations are evaluated, so the significance level is a (slightly conservative) 2 in 10000, i.e. p<0.02%. There is little merit in then considering the pair-wise tests of Block1 v Block2, Block 1 v Block3 etc. The key thing to have established is that there are natural changes in the nematode assemblage across the sandflat, so that removing this block factor from the test for treatments was worthwhile (the *n*MDS shows that a 1-way design in which block-to-block changes become part of the replicate variability would largely fail to pull out the treatment effect). Individual block differences are not of interest, but if they were, the *Pairwise Tests* table in ANOSIM1 shows that all blocks are well separated from each other (all pairwise R values are large). Note that none of these pairwise comparisons has enough replicates (and thus permutations) to allow a sensible significance test. In all cases the observed configuration is the most extreme permutation, best separating the two blocks, but with only 9 permutations that gives significance of only p = 11.1%. It is not logical to conclude that there are no differences between any pair of blocks when the global test has just shown that there are massive and highly significant differences amongst all four blocks! As remarked for 1-way ANOSIM, the focus should be on showing significance of the global R (otherwise pairwise comparisons should not be pursued) and then the pairwise R values themselves, to see where the large effects are (here, between all pairs).



2-way crossed ANOSIM (Danish sediment data)

For an example of a 2-way crossed ANOSIM test in a very different context, save and close the above workspace and return to the particle size distributions from Danish sediments introduced at the very end of Section 4 – the workspace Denmark ws in C:\Examples v7\Denmark PSA, with particle size frequency data (over 6 size-categories) in Denmark PSA histogram, for 3 sites (A, B, C) crossed with 2 depths (2m and 5m), and 5 replicate samples from each combination of site and depth. A distance measure such as Euclidean (or Manhattan or Maximum distance, see Section 5) is appropriate for defining the resemblance between two distribution curves.

As suggested earlier, apply this to the (smoother) cumulative frequencies from **Pre-treatment> Cumulate Samples>**(Variable Order•As worksheet), and on the resulting resemblance matrix run **Analyse>ANOSIM**, with the 2-way crossed option for factors Site and Depth (in some cases, the latter might be considered ordered, but we have only 2 depths here so the distinction is irrelevant; Site is clearly unordered). The results show perfect separation of the depths (global average R = 1, p<0.1%) and strong separation of the sites (global average R = 0.80, p<0.1%). The large number of possible permutations means that these p values could be made almost arbitrarily small, as is clear from the null distribution histograms. Pairwise site tests show, however, that B and C are not well separated (average R = 0.36, though still larger than all but 2 of the 999 permutations considered from the full set of 15876, thus p<0.3%), as is also seen in an *n*MDS plot on the Euclidean distance matrix (or a PCA, see Section 12). Close this workspace – it will not be needed again.



(Phuket coral reefs)

The study of coral reef assemblages at Cape Panwa, Phuket, Thailand – introduced near the end of Section 8 – measured area cover of corals on twelve 10m line samples taken perpendicularly to an onshore to offshore transect (A), over a time series of years. The previous workspace, Phuket ws, contained the data only for 1983-87, and 1983 was selected to visualise community change along the A transect, seen in an *n*MDS. Here, you should open into that workspace (or a clear one) data from the 7 middle years, Phuket coral cover 88-97 in C:\Examples v7\Phuket corals. Factors Year (88, 91, 92, 93, 94, 95, 97) and Position on the A transect (1-12) form a 2-way crossed design, because each position is examined (*plotless line sample*) in each year, but there is no replication.

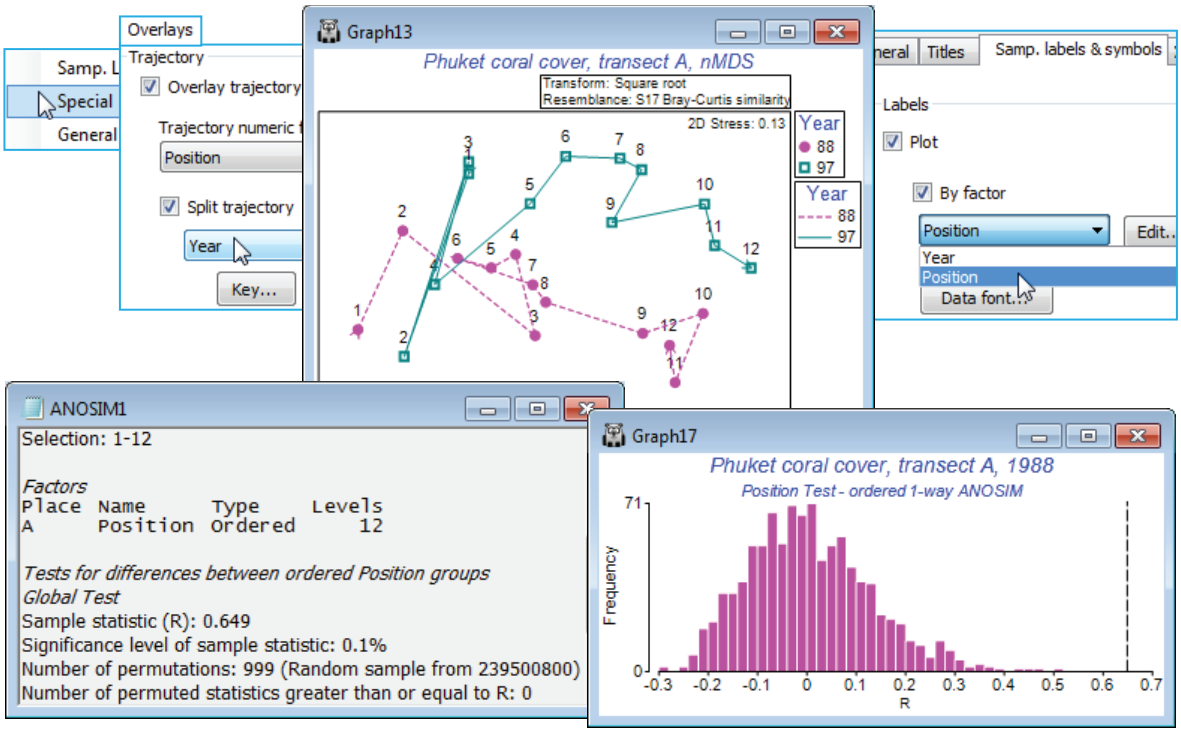
1-way ordered without replication

v7

In the unordered 1-way design, replication is essential for any sort of test (otherwise how can you tell whether single samples from groups A, B, C, … are from the same or different communities? – there are no within-group rank dissimilarities to compare with among-group ones). For the ordered 1-way design, however, the test statistic R^O can still be constructed – see the explanation in CiMC Chapter 6 under ANOSIM for ordered factors, where the statistic for the unreplicated design is designated R^{Os}, for *ordered single*, rather than R^{Oc}, for *ordered category* (though in both cases it is fundamentally the same slope statistic R^O from a regression of rank dissimilarities against modelled rank distances under the alternative hypothesis). A univariate analogue you may find it helpful to think about is testing whether differences in a variable *y* bear any relation to given values of *x*. If

you are not prepared to make any assumptions about the form of the relationship (the alternative hypothesis just says the values of *y* differ with those of *x* in some way unspecified) then you must have replicates at each *x* value in order to construct an (ANOVA-type) test. If, however, you set out to examine the alternative hypothesis that the relationship between *y* and *x* is linear, then there is a perfectly viable test without any replication of *x* levels, i.e. whether the slope of a linear regression of *y* on *x* is significantly different from zero. And you may choose that linear regression test even when there are replicates at each *x* level. This is actually a very precise analogue of the difference between ordered R^O (regression-type) and unordered R (ANOVA-type) ANOSIM tests.

So, for the Phuket coral cover 88-97 data, take a square-root transform and Bray-Curtis similarities, selecting from the latter the first and last years 88 and 97 (i.e. 24 samples, the 12 transect positions in each year) and reproduce the *n*MDS plot seen in Fig. 6.14 of CiMC – with separate trajectories over transects for each year by taking **Graph>Special>Overlays>**(✓Overlay trajectory Position)>(✓Split trajectory Year) and on **Samp. labels & symbols**, (Labels✓Plot)>(✓By factor Position). It is scarcely necessary to test the null hypothesis of no *Position* effect for each of these years but a 1-way ordered test (without replicates) can be carried out by selecting each year in turn, and **Analyse >ANOSIM>**(Model•One-Way - A)>(Factors A: Position)>(Type Ordered) gives R^{Os} = 0.65 and 0.73 respectively (both p<0.1%).



2-way crossed ordered test

The test for an ordered factor (A) in the 2-way crossed design parallels the construction seen earlier for the 2-way (unordered) crossed case, in that the 1-way R^O statistic is calculated separately for each level of the other factor (B) and those R^O values averaged to give the 2-way test statistic. This is compared with its null distribution calculated under the same constrained permutation procedure as for the previous 2-way crossed case – A labels are permuted only within the levels of B. The difference here, again, is that this is a perfectly viable test when there is no replication within the cells of the 2-way layout, provided there are enough ordered steps (*a*) in factor A or levels (*b*) of factor B to generate sufficient permutations, $(a!/2)^b$, for a sensible test. This number scales up very rapidly, so even a fairly minimal design will give some sort of test, e.g. *a*=4 transect sites sampled *b*=2 times gives 144 permutations and (at best) a p<1% level test for the presence of site ordering. The 1-way test (*b*=1) requires at least *a*=5 ordered steps, to give 60 permutations for a p<2% test.

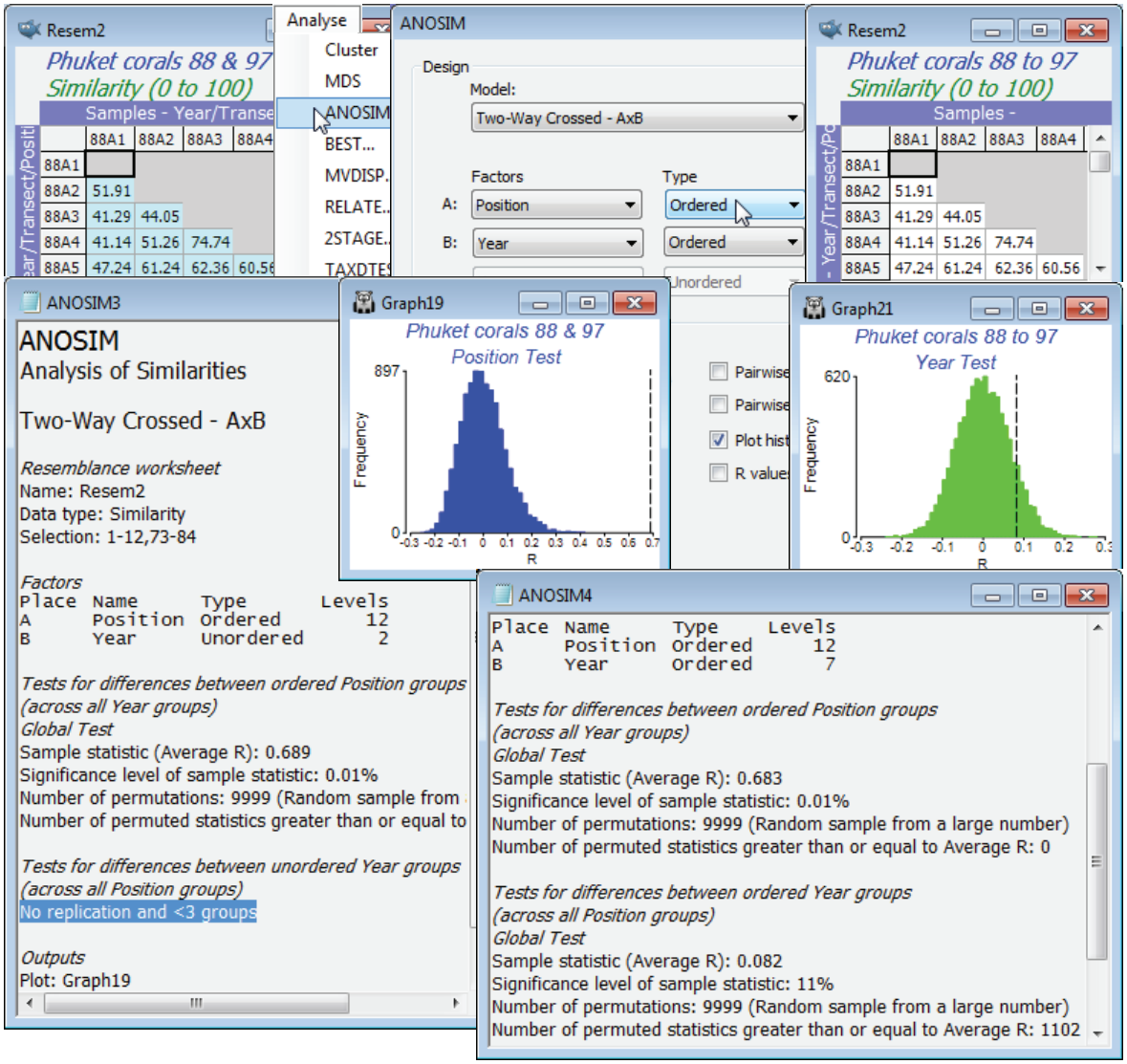
Run **Analyse>ANOSIM>**(Model•Two-Way Crossed - AxB)>(Factors A: Position Ordered & B: Year Unordered) on same resemblance selection as above, of the two years 88 and 97, together. This will, of course, produce a massively significant *Position* effect, with average R^O = 0.69, and with (Max permutations: 9999) this is still off the top of the null distribution, p<0.01% (or, as a probability, P<0.0001). It is naturally a very powerful test, with 6×10^{16} possible permutations.

It did not matter in this case whether the Year factor was defined as Unordered or Ordered, since there were only two years. The test for *Year*, removing the effect of *Position* by comparing years only within each of the 12 levels for *Position*, is doomed to failure, unsurprisingly. There are no replicates on which to base such a test (applying the above formula for an ordered test, a=2 so $a!/2 = 1$ and, whilst b=12 is large, powering up 1 still gives 1, i.e. there is only one permutation which is the observed configuration of the labels!). ANOSIM simply says *No replication and <3 groups*.

v7

However, if we were to take off the selection, so reintroduce the full set of 7 years, and specify that both factors are ordered then there are ample steps in both the spatial gradient of 12 points and a temporal time trend of 7 points for an ordered test of either factor, removing the effect of the other. The *Position* test now gives a very similar $R^O = 0.69$ as found for the two years alone but the *Year* test returns $R^O = 0.08$, with about 1100 of the 9999 permutations created under the null hypothesis giving larger R^O values than this ($p<11\%$), a non-significant result. (Incidentally, note it is always true that a test of factor A is completely unchanged by whether factor B is assumed ordered or not).

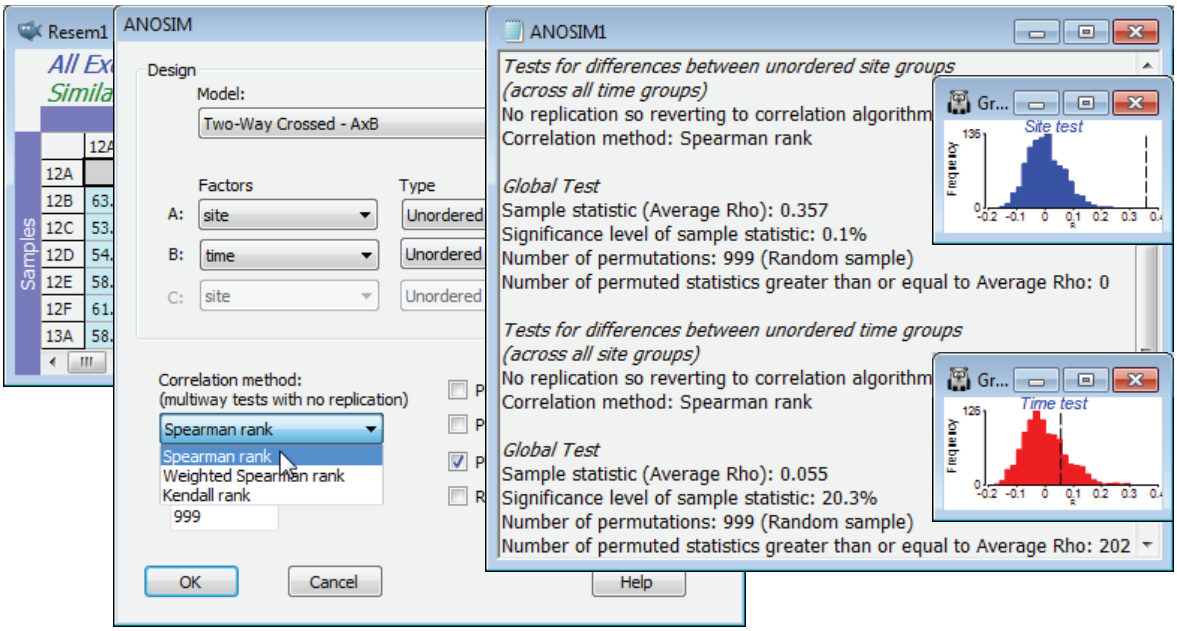
In fact, whilst the original study postulated serial change in coral communities along the onshore-offshore transect, so that an ordered test for the *Position* factor seems very appropriate, it is not so clear that it is relevant to test for a monotonic inter-annual trend – a drift of the community in time, ever further away from its original configuration. Local impacts in some years may be dominant, and the possibility that these have a differential effect on the transect gradient (an *interaction* of a type) suggests a very different approach, using the **Analyse>2STAGE** routine, which we shall return to for these data in Section 14. Within the ANOSIM routines however, the 2-way crossed layout for an unordered factor with no replication leaves few options for a non-parametric test, though sometimes helpful is a fall-back test (available in PRIMER since the early versions), which is next described for the Exe estuary nematode data. Save and close workspace Phuket ws.



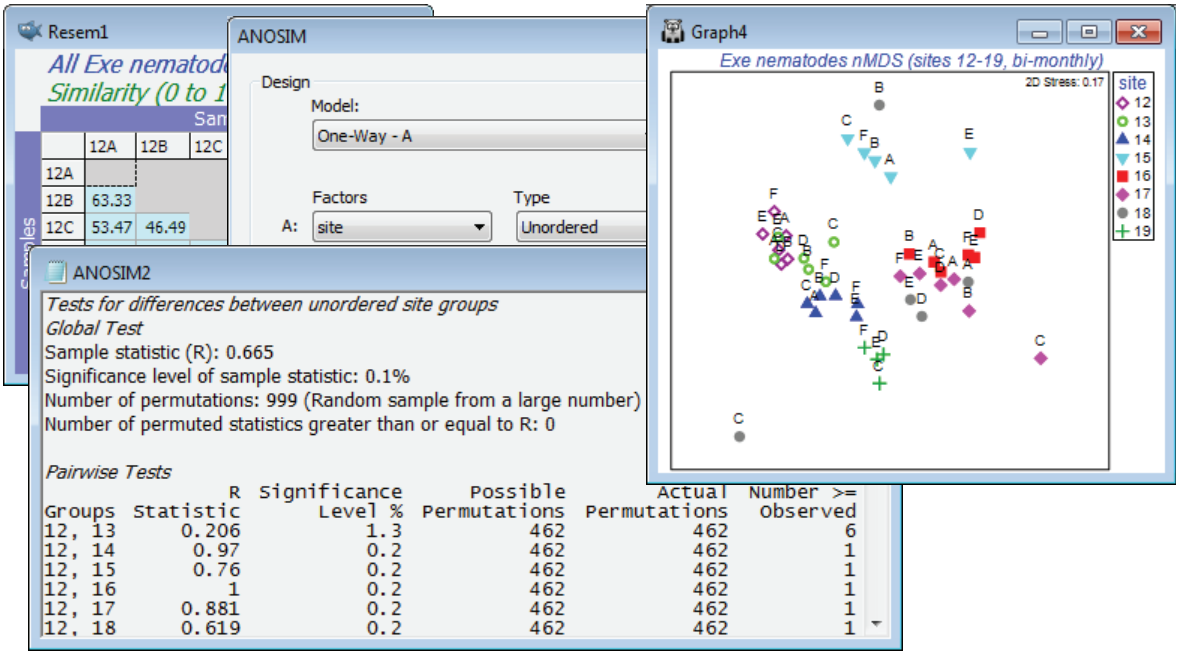
ANOSIM for 2-way crossed design with no replication (Exe study)

The 2-way crossed ANOSIM for an unordered factor, and with each combination of the two factors only having a single replicate, is covered in CiMC, Figs. 6.9 to 6.12, firstly for a treatment × block design and then for the example considered here of sites crossed with times. This is the inter-tidal Exe estuary nematode study, first seen at the start of Section 6 and used to demonstrate clustering and MDS, but here you should open the full data in a new workspace, i.e. the bi-monthly samples from the 19 sites, Exe nematodes bi-monthly in directory C:\Examples v7\Exe nematodes. It is the 6 seasonal samples, covering one year, which were averaged for each of the 19 sites in the earlier analysis of Exe nematode abundance. That there are clear site differences was obvious from the stark clustering of that data into 4 to 5 groups, but it is less clear whether there are site differences in the largest cluster, sites 12-19. So, as before, pre-treat the full data Exe nematodes bi-monthly with a 4th-root transformation and apply Bray-Curtis similarities, then select sites 12-19 (of course it does not matter whether you make the selection of these sites before or after calculating the transformations and similarities). The factors site and time are crossed because the *same* set of sites is returned to at each time. It is of interest here to test both, separately: are there differences among sites, removing the effect of times, and is there a seasonal effect, removing any site differences?

Analyse>ANOSIM>(Model:Two-Way Crossed - AxB)>(Factors A:site & B:time) both Unordered runs a different style of permutation procedure, testing for a site effect by asking whether there is evidence for commonality of the among-site pattern across the different times. For example, if the MDS plots of sites, displayed separately for each time (Fig. 6.12 in CiMC) show the sites grouping in the same way, that must imply there are site differences. To put it the other way round, under the null hypothesis that there are no site differences, the separate MDS plots for each time will have no common pattern and look like random rearrangements of each other. In fact the test operates, as with other ANOSIM tests, on the underlying resemblance matrix (ranks) rather than the MDS plots, and calculates an average of all pairwise correlations (ρ_{av}) between the among-site resemblance matrices for each time; ρ_{av} will be near zero if there are no site effects. (The idea of this correlation ρ between two triangular matrices, a type of non-parametric Mantel statistic, is at the core of most of Sections 13 & 14, and is discussed extensively in Chapters 6, 11, 15 & 16 of CiMC). ANOSIM then recomputes this ρ_{av} statistic for random permutations of the site labels at each time (since for the null hypothesis there are no site differences), to obtain a null permutation distribution for ρ_{av}, and thus a significance test. CiMC, equations (11.3) to (11.4), gives details of the choices offered by the ANOSIM dialog, of rank correlation coefficient ρ to calculate between pairs of resemblance matrices – Spearman rank is the best known and the default. Note that the routine automatically copes with a small number of missing samples (here caused by weather and/or tidal states at one or two sites on one or two occasions) because for each pair of times it can drop the sites which are not found in both configurations (called pairwise deletion of missing values), without having to drop those sites from the whole matrix (called listwise deletion). A satisfactory test does, however, need a decent number of shared sites available for all pairs of times (and for the test to have any power at all, some interactions must be small, otherwise no commonality is detectable – Chapter 6, CiMC).



The results show a significant site effect (ρ_{av} = 0.36, p<0.1%) but not evidence for a seasonal effect (ρ_{av} = 0.06, p≈20%, i.e. the relationships amongst times do not show commonality over all sites, or over sufficiently many of them to depart from random re-arrangement). The latter finding is not so surprising in a climatically mild region, given that generation times of meiofauna are measured in weeks. We might therefore be justified in strengthening the testing procedure for sites by running a 1-way ANOSIM on site, using the full set of 44 samples from sites 12-19, i.e. treating the different times as replicates. Now we can obtain tests between pairs of sites. Such pairwise comparisons are not available with the 2-way crossed analysis (without replicates) for obvious reasons – one cannot ask about commonality of pattern across 6 MDS plots, if each consists only of 2 sites, i.e. a single similarity value! From the 1-way ANOSIM results and the MDS of all samples, it is clear that most sites have significantly different and well-separated assemblages – with the exception of site 18, which is species-poor and has widely scattered replicates (time points) on the MDS, and site 12 *vs* 13 and site 16 *vs* 17, with low R values of 0.21 and 0.17 respectively. Close the workspace.



2-way nested ANOSIM (Calafuria macroalgae)

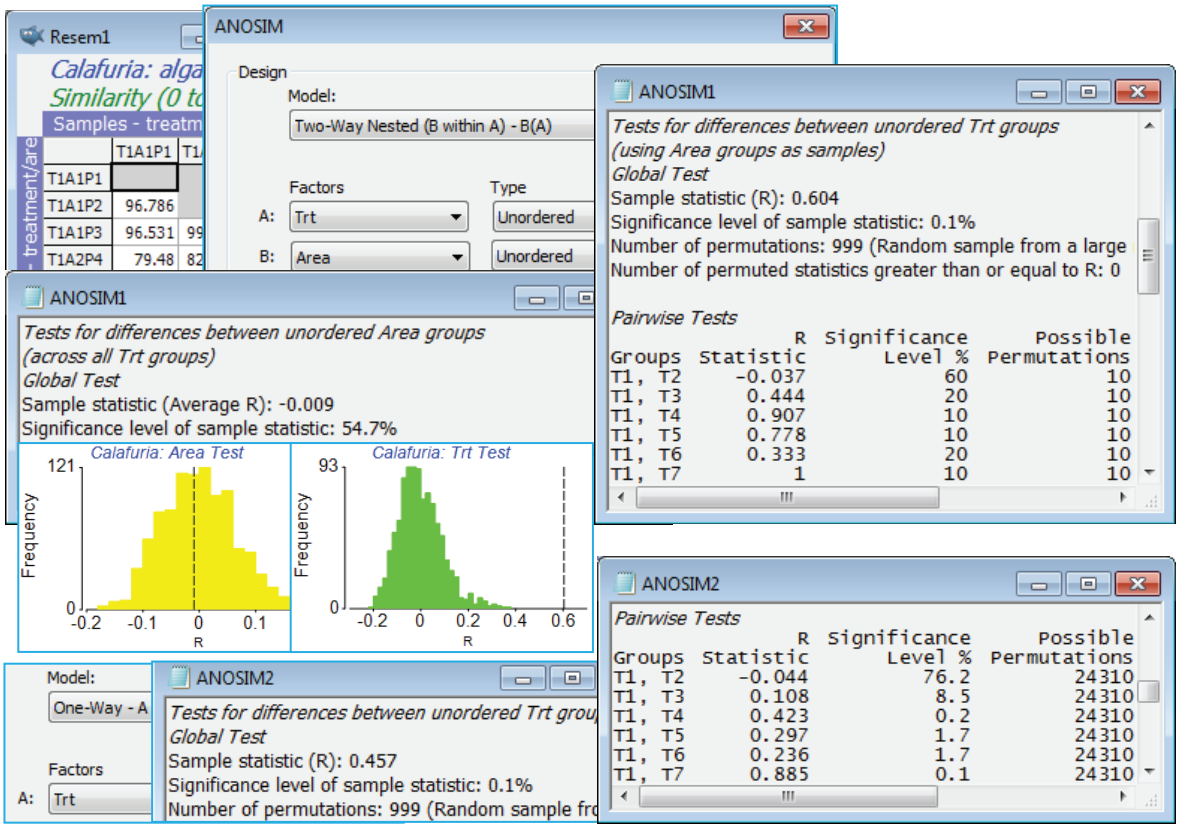
Subtidal rocky reefs, at c10m depth, at the Calafuria station in the Ligurian Sea, N Italy, were the subject of a clearance and recovery experiment by Airoldi L 2000 *Mar Ecol Prog Ser* 195: 81-92 (see also Clarke KR, Somerfield PJ, Airoldi L, Warwick RM 2006, *J Exp Mar Biol Ecol* 338: 179-192; both sources analyse a wider set of data than considered here). For 8 different times between October 1995 and September 1996 (factor A: named *treatment*, with levels 1 to 8), rock patches were cleared from three randomly chosen areas (factor B: *area* with levels 1 to 24, the three areas differing for each 'treatment', naturally). Three randomly chosen *plots* from each area (replicates) were then examined at the end of one year of recolonisation, and % area cover recorded of nine macroalgal taxonomic categories. The design is therefore a *2-way nested* layout, with factor A (Trt) at the top level and B (Area) nested within A, denoted B(A) – replicates can be thought of as nested within B (replicates are always nested). Note that when defining the factor levels, PRIMER does not mind if you code the area levels as 1 to 24, or 1 to 3 repeatedly for each of the 8 treatments. If you use the latter, when 2-way nested is selected under **Analyse>ANOSIM**, and *area* is specified as nested, the routine will know that there is nothing in common between area 1, treatment 1 and area 1, treatment 2. But it may help you to code the area levels as 1 to 24, because the fact that area is nested, not crossed, with treatment will then be clear.

Open Calafuria algal cover from C:\Examples v7\Calafuria algae, and examine its factors with **Edit >Factors**. A strong transform is necessary to prevent the taxon category *Algal turf* from completely dominating, so transform with fourth root and calculate Bray-Curtis similarities between samples. The primary interest is in whether there are differences in recolonised macroalgal communities a year after clearance, depending on the time of year at which the clearance took place, i.e. a test of the treatment factor. But it is important to choose the correct replication level for this test. This is usually (some would argue, always) the level of variability immediately below treatment in the

hierarchy, namely the areas not the plots, which are a level further down. So, one possibility is simply to average the three plots within each area and carry out one-way ANOSIM on the Trt factor (with 3 replicate areas per treatment). But what if there is absolutely no area effect, i.e. plots in different areas are no more dissimilar from each other than plots in the same area? Then it would seem reasonable to take plots as the replication level for testing treatment effects, and the much greater number of replicates will improve the sensitivity of that test.

On the resemblances, take **Analyse>ANOSIM>**(Model: Two-Way Nested (B within A) - B(A))> (Factors A: Trt) & (B: Area), both of which are Unordered (the areas are randomly chosen from the region at each starting time, Trt, and though these starting times run sequentially through a year, a serial pattern for the algal recovery state a year later would not be expected – if anything it will be *cyclic* with the seasonality, see Section 14 for such tests). The routine then carries out two tests. Firstly, it tests the null hypothesis that there is no area effect. Nothing is assumed about treatment effects; these may or may not be present but need to be removed, in exactly the same way as for the 2-way crossed ANOSIM (i.e. R values contrasting among- and within-area rank dissimilarities are calculated separately for each treatment and averaged; permutations are constrained to shuffle labels among plots only over areas within a treatment, not across treatments, etc). Secondly, the routine then always presumes that an area effect is present, so tests the treatments by averaging the plots within areas, thus using areas as the replication level for this test, by a 1-way ANOSIM. (In fact, the averaging is done on the rank dissimilarity matrix, which is then re-ranked for the 1-way ANOSIM – see CiMC). If there is demonstrably no area effect at all, so the test can use all 9 plots as replicates of a treatment, this needs a separate run of 1-way ANOSIM, ignoring the area factor.

The 2-way nested test for areas here gives average R = –0.01 (p≈56%) and this near-zero R implies absolutely no suggestion of an area effect, making the 2-way test for treatments (averaging up to area level) unnecessarily conservative. It still gives a strongly significant global R of 0.60 but the conservatism is seen in the pairwise table, where comparisons are based on only 10 permutations (3 areas for each treatment). If, as is justified here, we ignore the area effect, the 1-way ANOSIM (9 plots per treatment) gives a pairwise table with 24,310 permutations for each comparison and thus clear inferences, e.g. T7 and T8 differ the most strongly from other times, with most R values in excess of 0.8, whereas pairs not involving these two times generally give R<0.4. If the initial test for area had given R>0 however, and certainly if it had been significantly so, on what will usually be a powerful test (many permutations), then it would not be justifiable to ignore the area effect and use plots as replicates: this would be non-conservative (*pseudo-replication*). Close the workspace.



3-way crossed ANOSIM (King Wrasse diets)

A dietary study of W Australian fish concerns composition by the volume of taxa (21 broad dietary categories: gastropods, bivalves, annelids, etc) in the foregut of King Wrasse from one of 4 length-classes, caught in 3 locations in 2 periods of the year and 2 replicate times of sampling within each period (each replicate is a similar-sized pool of gut content of fish in each length-class). The data is Wrasse gut composition in C:\Examples v7\Wrasse diets. More detail is given in Chapter 6, CiMC; the analysis is from a wider study by Lek E *et al* 2011, *J Fish Biol* 78: 1913-1943.

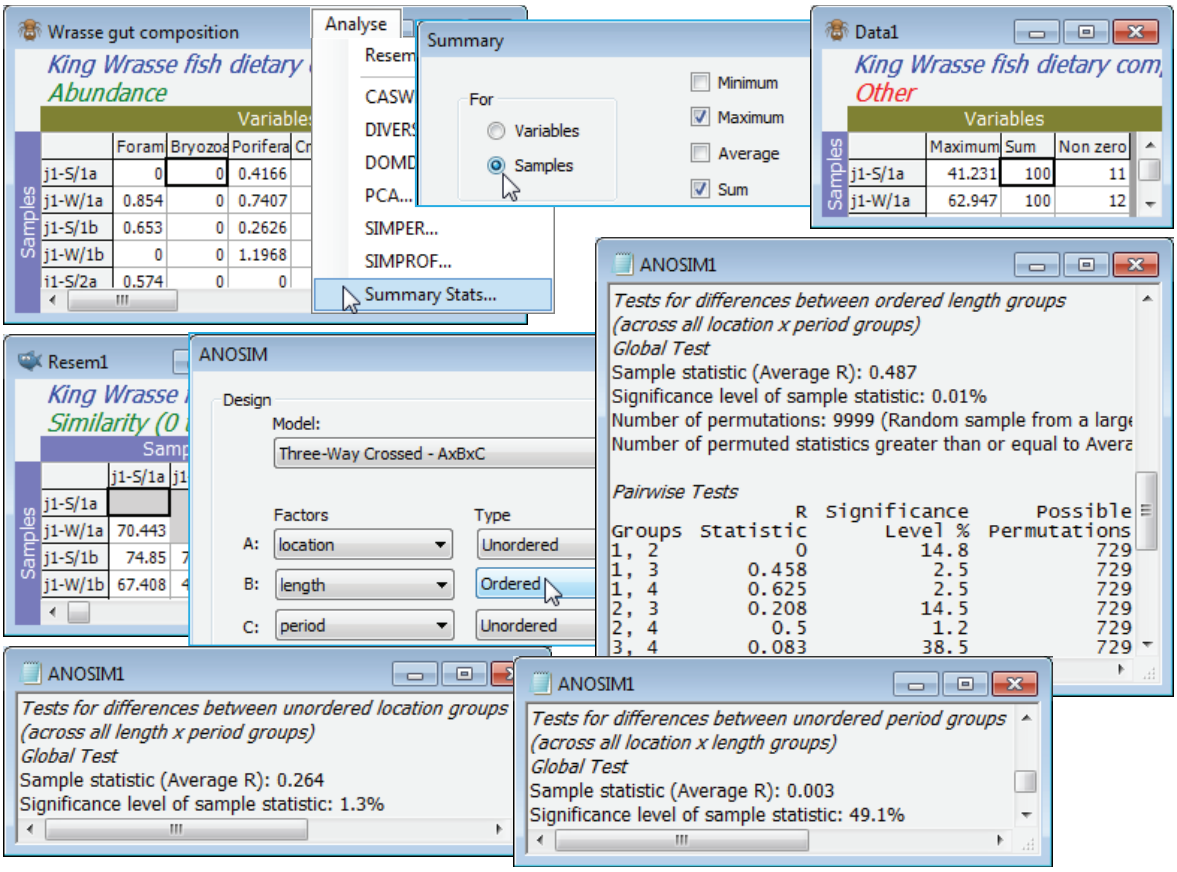
v7 This is an example of a 3-way fully crossed design with replication, A×B×C, with A: location, B: length and C: period. A test of the null hypothesis of no effect, of each of the factors in turn, simply uses the earlier 2-way crossed design, e.g. with first factor A and second the *flattened* B×C factor. The latter places all combinations of the levels of B and C in a single factor (using **Edit>Factors> Combine** and placing the B and C factor names in the Include box). The 2-way test for factor A now therefore constructs a 1-way ANOSIM R statistic for each combination of levels of B and C, and averages those, testing this against permutations constrained to stay within the B×C levels. The same procedure is followed for each factor, so B is tested having removed the effect of A×C, and C

v7 is tested have removed any effect of both A and B (or their interaction). So, whilst this could all be carried out by three separate runs of 2-way crossed ANOSIM, it is more conveniently (in PRIMER 7) performed by specifying the 3-way crossed design A×B×C. The resulting three average R values can then validly be compared, to determine the relative overall magnitude of the three effects.

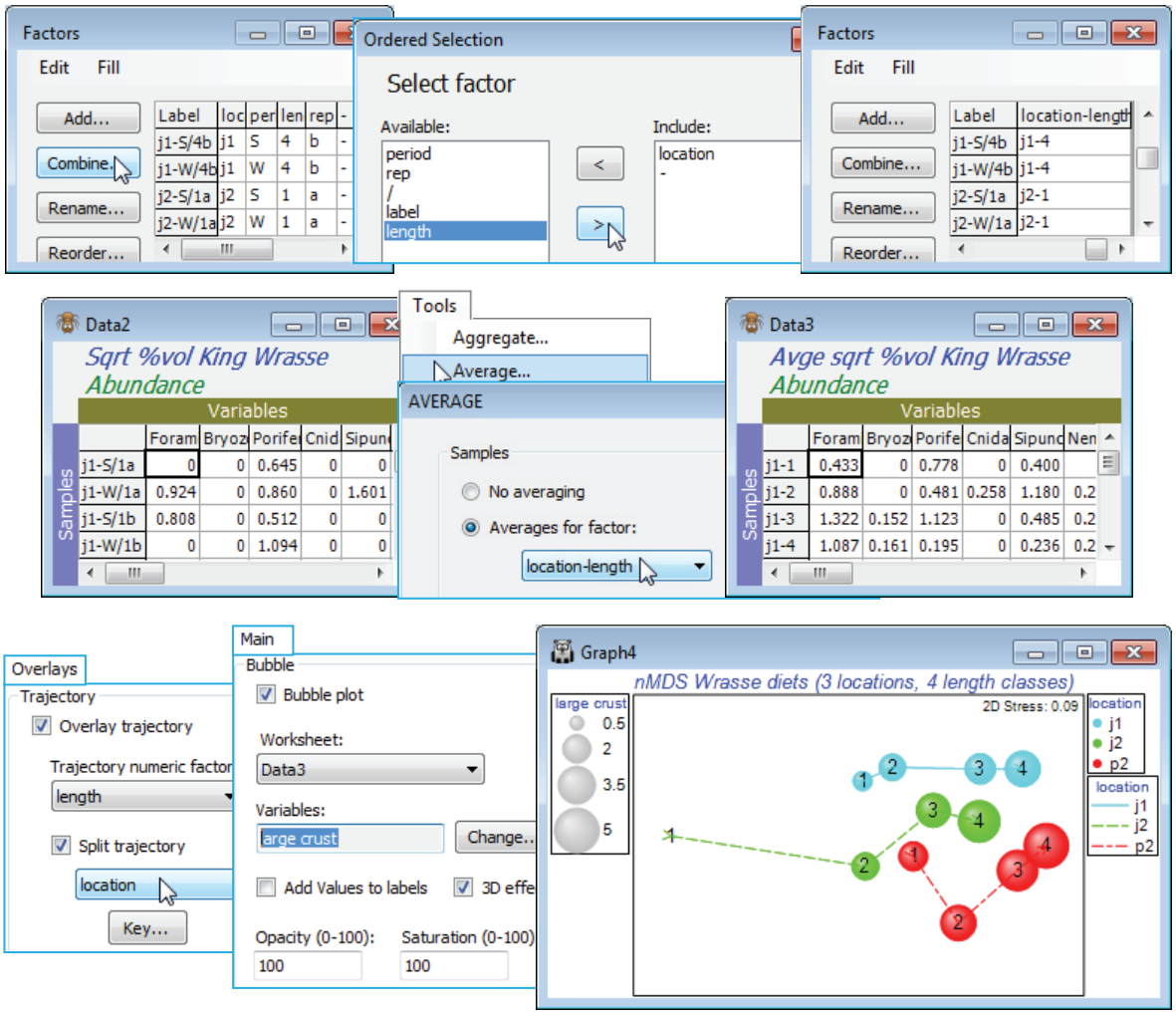
Here, the 3 levels of the *location* factor are not ordered, and there are only 2 *period*s (so whether treated as ordered or not is immaterial), but the 4 *length* classes of the predator wrasse are clearly ordered – the expectation is that, if the diet changes at all, it will do so in progressive fashion as the

v7 fish matures. So, check that the samples are already standardised to add to 100% across the dietary categories by **Analyse>Summary Stats>**(For•Samples) & (✓Sum) – you may want to leave other boxes checked also – and square-root transform then compute Bray-Curtis similarities, and input to

v7 **Analyse>ANOSIM>**(Model: Three-Way Crossed - AxBxC)>(Factors A: location) & (B: length) & (C: period), with B Ordered, specifying (Max permutations: 9999). The average R values show that B (0.49, p<0.01%) is the largest effect, then A (0.26, p<1.5%), but that C has no effect at all (0.00). The pairwise average R values for the length-class effect show the increasing differences in diet with difference in fish lengths, exactly as would be expected for an ordered factor (you may wish to re-run the analysis for an unordered B factor and note how this reduces its global average R value).



Given the clear absence of a period effect, a useful summary would be to average the (transformed) data over both the replicates and the periods (by **Tools>Average** and supply the combined factor for A×B), recalculate the similarities and enter *n*MDS. The plot exemplifies a split trajectory, with **Special>Overlays** for length, splitting by location; you might also like to recreate the bubble plot of Fig. 6.15 of CiMC by adding a dietary component. Save and close the workspace (Wrasse ws).



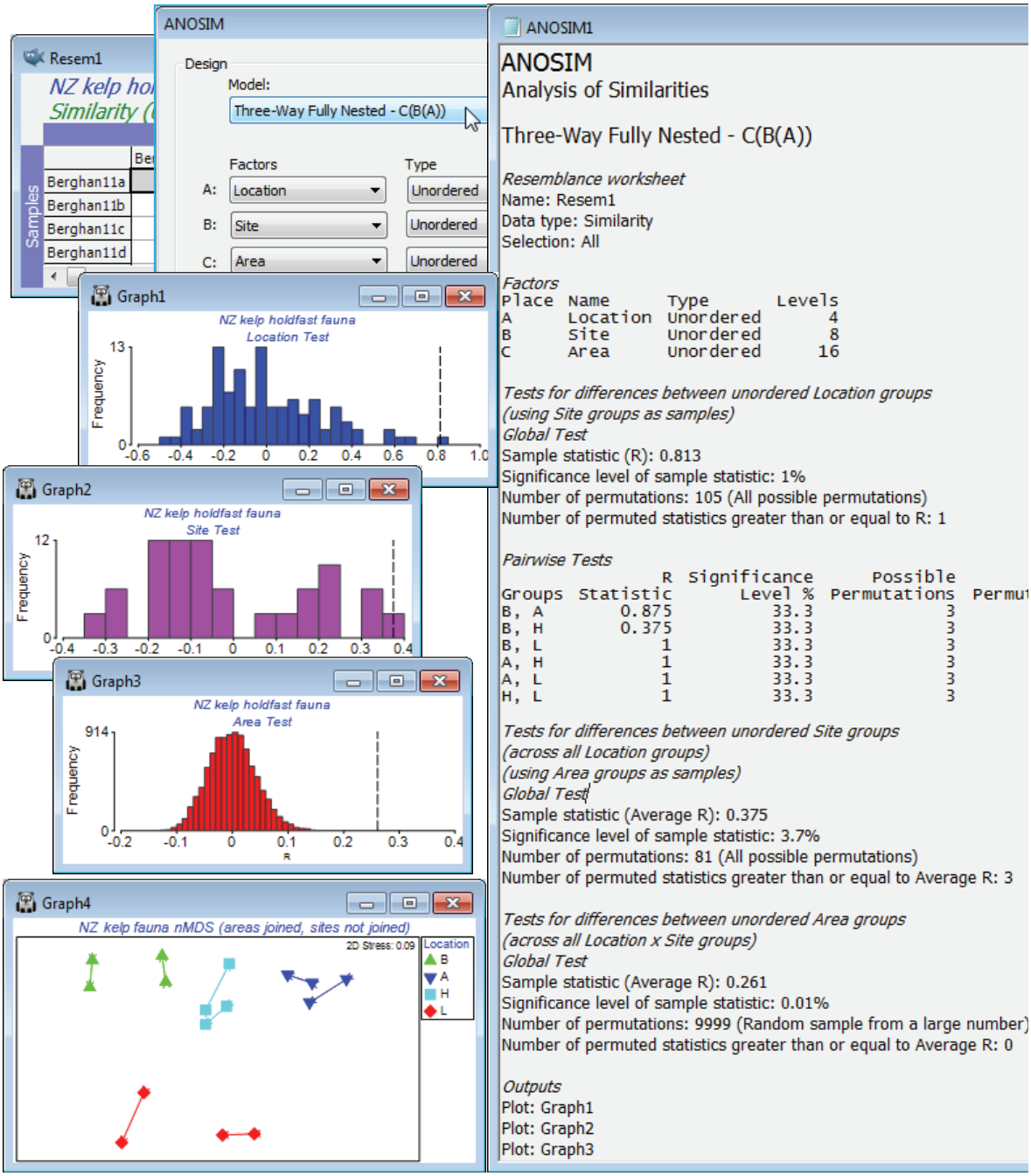
3-way fully nested design (NZ holdfast fauna)

The 3-way fully nested design has factor C at the lowest level, nested in B at the mid level, which itself is nested in A at the top level, denoted C(B(A)). Factors can again be ordered or not, and the routine is essentially a repeated application of the 2-way nested design above – the first test, for C, is carried out simultaneously within the strata of all B levels (for every A level), the replicates in C levels are then averaged (in the same way as for the 2-way test, by averaging appropriate similarity ranks) and the test for B and A are now exactly that of the 2-way nested B(A) design. If replicates at the C level are not felt to be particularly reliable as snapshots of the community (each is species-poor, though pooled they give a fair representation of species presences at each level of C), it may be more efficient for the tests of B and A to pool or average the replicates in the data matrix, rather than the (rank) similarities, and run a 2-way nested B(A) ANOSIM with C levels as replicates.

An example can be drawn from a data set of Marti Anderson and colleagues (Anderson *et al* 2005, *J Exp Mar Biol Ecol* 320: 33-56) distributed with the PERMANOVA+ add-on software, analysed in detail in the PERMANOVA+ manual (Anderson *et al* 2008) but which is also now to be found in C:\Examples v7\NZ holdfast fauna, as data file NZ holdfast fauna abundance. Chapter 6, CiMC gives the three-way nested ANOSIM tests for these data, see Figs 6.16 & 6.17. The macrofauna found in kelp holdfasts was sampled at 4 northern New Zealand *Locations* (A), with 2 *Sites* (B) per location, sampling 2 *Areas* (C) at each site, with 5 replicate holdfasts at each area. Clearly, *Areas* are nested in *Sites*, which are nested in *Locations*, C(B(A)). With only 2 sites per location and 2 areas per site, neither factor can be considered ordered, and there is also no case for considering the top-level locations ordered.

v7 After square-root transformation and with Bray-Curtis similarities, **Analyse>ANOSIM>**(Model: Three-Way Fully Nested - C(B(A)))>(Factors A: Location) & (B: Site) & (C: Area), all Unordered, and (Max permutations: 9999). The resulting test statistics: R = 0.81 (p≈1%) for the location test, and average R = 0.38 (p≈1%) for sites and 0.26 for areas (p<0.01%), are again directly comparable with each other as measures of the extent to which stepping up the spatial level (replicates to areas, areas to sites, sites to locations) results in additional community differences – the largest effects are clearly at the location level. (Note the importance of interpreting the R values not the p values – the latter are always hijacked by the differences in number of permutations, here respectively 105, 81 and infinite, effectively, so that the smallest R value is actually the most significant!). Now produce a summary of these community differences at the different levels of the design, by averaging the square-rooted abundances over the replicate level (since the areas have all, sensibly, been given a different number, irrespective of the site or location, **Tools>Average** for factor Area will achieve this), then recalculating similarities and running *n*MDS. By careful use of symbol key changes, the

v7 means plot of Fig. 6.17, CiMC can be produced: plot symbols by Location; overlay trajectories by Area, split by Site; match up the line colours in pairs with those of the *Locations* and make all the lines continuous by clicking on the *Site* line key next to the plot; finally remove the *Site* line key by unchecking the (✔Plot key) box for Site on the **Key** tab, accessed through (say) **General** – easy!



3-way crossed /nested design (Tees Bay macrofauna)

v7

The two other possible 3-way designs can be written C(A×B) and B×C(A). The first is straightforward: an example might be of locations (A) each containing the same set of habitat types (B), and within each combination of habitat and location a number of sites (C) are randomly chosen, with replicates taken at each site. The ANOSIM tests are again effectively 2-way cases, with A and B flattened by combining them into a single factor (AB), so the design for testing C is the nested case C(AB), and the tests for A and B using the crossed 2-way A×B design, with the levels of C as 'replicates' (averaging over the original replicates is, as usual, on the rank similarities). The 3-way C(A×B) choice handles all this automatically, naturally, and again factors can be ordered or not.

The remaining possibility B×C(A), in which B is crossed with all levels of C, the latter nested in A, is more complex and for some tests requires variations of the 2-way test statistics (fundamentally still either average R, R^O or ρ constructions, however) and a modified permutation procedure in which whole sets of sample labels are permuted together. This is discussed extensively in CiMC, towards the end of Chapter 6 and in Table 6.4, and will not therefore all be repeated here. We will just use the same example as in CiMC to illustrate setting up and interpreting these tests. This is the Tees Bay macrofauna data introduced in Section 8 as an example of time-series trajectories in ordinations (see also Fig. 6.17 in CiMC), with workspace Tees ws and data Tees macrobenthic abundance in C:\Examples v7\Tees macrobenthos. Factors are Area (A), Year (B) and Site (C), with the 4 areas of Tees Bay (1-4) each containing two sites (a, b), each of which was resampled

v7

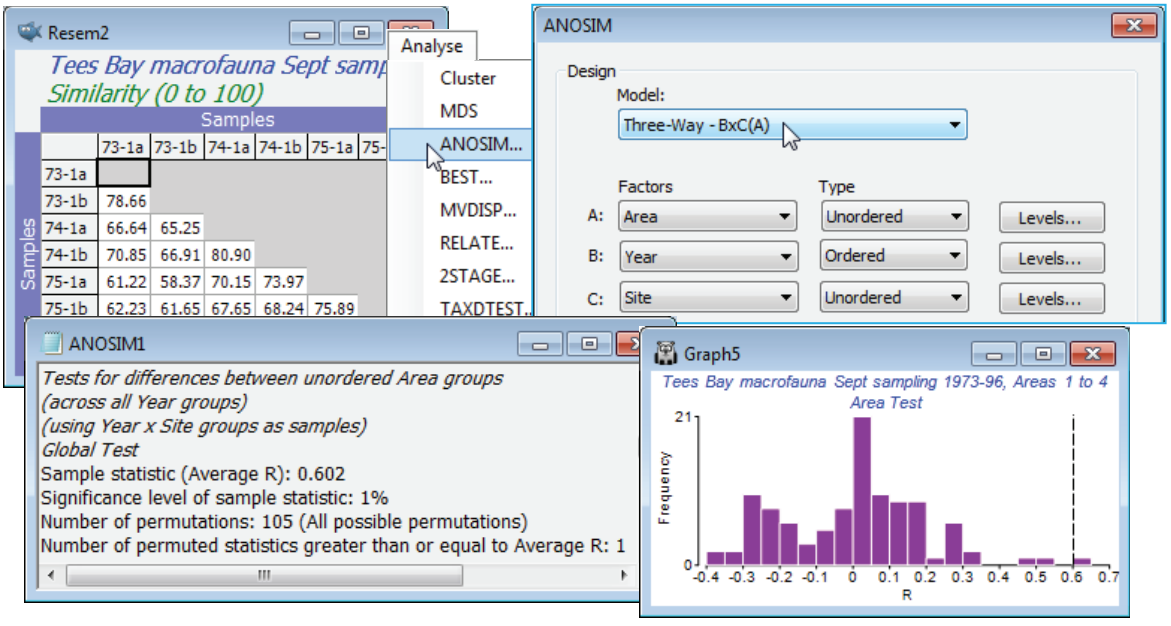
every September over the period 1973-96. Clearly the *Sites* are nested in *Areas* but *Sites* and *Years* are crossed (all sites sampled in all years), hence the design is B×C(A). The data in this file has no replication at each site by time combination, the original multiple grab samples collected from a single sampling visit being regarded (perhaps a little harshly!) as *pseudo-replication* in time, and possibly even space, for September sampling of this community in a particular site and year – and thus the sample identifications from the multiple grabs (raw data) were averaged.

Areas are along a NW-SE transect of the coast but cannot be considered Ordered, since the mouth of the Tees estuary intervenes – see map in Fig. 6.17 of CiMC. [In fact, if the tests are done under the assumption that the areas are ordered 1 to 4, there is a failure to detect an area effect at all. This is a good example of the dangers of specifying the alternative hypothesis too narrowly – there is no power to detect an effect which does not conform to that alternative. Here the central areas 2 & 3 are different than the surrounding areas 1 & 4, probably because they are influenced by the Tees

v7

estuary mouth in the mid-Bay region]. *Years*, however, could be considered Ordered because there is interest in whether the communities show a (climate-change driven?) yearly trend, but could be entered as Unordered instead, if a simple serial drift is considered too narrow an alternative.

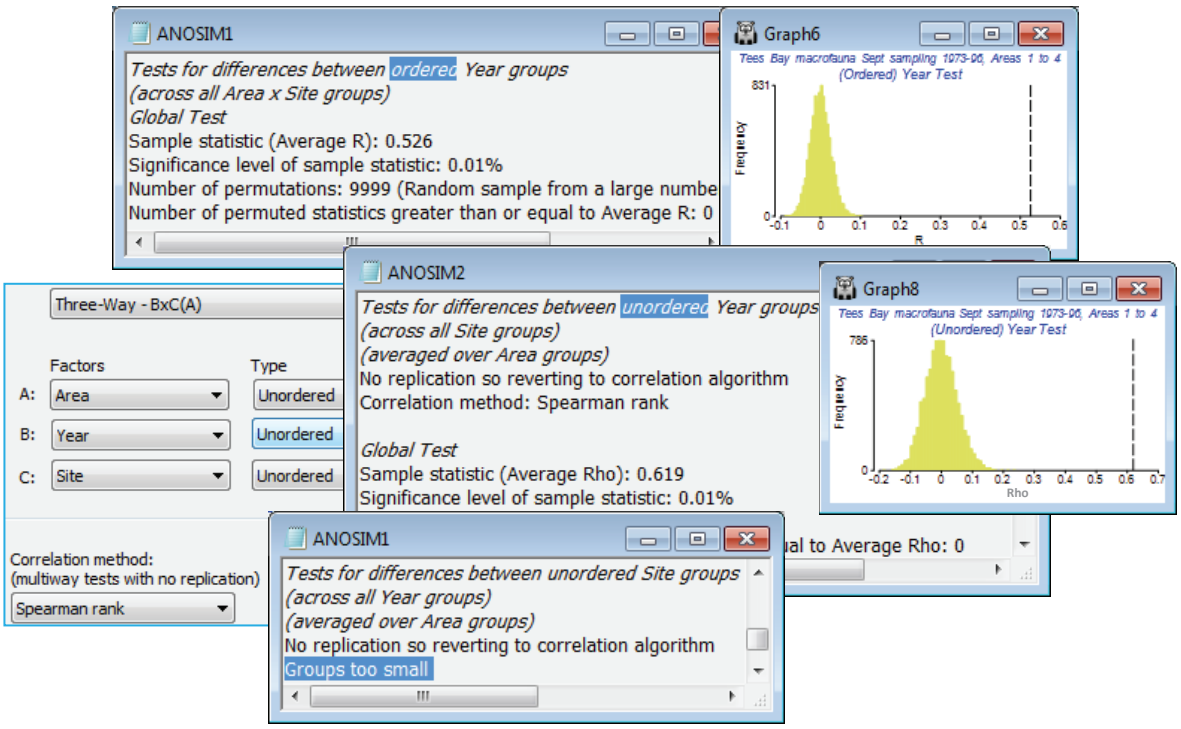
On 4th-root transformed data with Bray-Curtis similarities, **Analyse>ANOSIM>**(Model: Three-Way - BxC(A))>(Factors A: Area) & (B: Year) & (C: Site), with B Ordered, and then repeat the run with B Unordered, so the correlation method is needed: (Correlation method: Spearman rank).



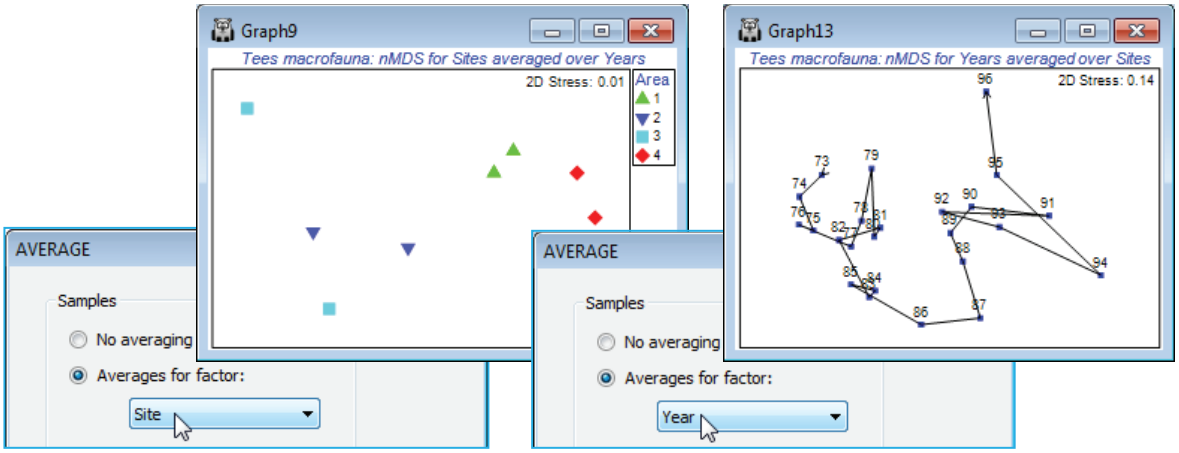
The (unordered) *Area* test has average R of 0.602 and this is the most extreme value in the 105 permutations (p<1%) – the permutation procedure here, a novel one for PRIMER, permutes the site labels within the areas but carries the full set of years with it, intact, reflecting the fact that the same sites are returned to each year, rather than randomly selecting two sites from each area at each time.

The ordered *Year* test gives an average R^O of 0.526 which the null histogram shows to be as highly significant as we wish to make it. The unordered *Year* test needs to use a matching statistic (ρ) of the year pattern over sites or areas, as indirect evidence of a time effect from its spatial conformity. The average ρ of 0.619 is also very highly significant. The ρ value cannot be compared with the R^O statistic, though both vary in the range (roughly) 0 to (definitely) 1, but the interpretation is clear – there is a broadly consistent time pattern across the region and it is fairly strongly serial.

A test of the third factor, *Site*, is clearly not possible. There are no replicates below this level to utilise, so ANOSIM tries instead the indirect route of matching the among-*Site* pattern over the set of *Years* – but with only two sites per area this must fail (giving a *Groups too small* statement in the results window). It has been important to retain *Site* as a factor however (rather than regarding it as a replicate) because the same sites are returned to each year and, with this 3-way procedure, the tests for *Area* and *Year* are designed to utilise that information in a more justifiable test.



The test results again justify a summary in terms of *n*MDS means plots. The area effect is clearly seen in an ordination using (transformed) data averaged over the years for each site, and the time series of years, now with (transformed) data averaged over all sites, certainly indicates a strong serial change (the same time series for the 4 area levels separately is seen in Fig. 6.17 of CiMC).



10. *Wizards* & species analyses (*Basic MVA, Coherence plots, Matrix display, SIMPER)*

Basic multivariate analysis wizard

v7

The three **Wizards** menu items carry out sequences of routines, all of which can be run separately but which it is either convenient or instructional to have bundled up in this way, at least until you are confident about the steps involved and can dispense with such a prescriptive (and proscriptive) approach. All three menu items are run with a data matrix, not a resemblance matrix, as the active sheet, and usually – though not exclusively – prior to any pre-treatment (they incorporate a limited choice of such options). If you are a novice user and, having opened a data sheet from Excel (with the help of the Excel File Wizard), you have little idea where to start, then the **Wizards>Basic multivariate analysis** is a simple instructional tool which leads you through the most commonly used steps in a multivariate analysis within PRIMER.

If the data has been created as of type Abundance (or Biomass) – this can be checked or changed with **Edit>Properties** – you are given options to standardise samples (the default is not to do this) and the usual choices of transformation (square root is the default), see Section 4, before Bray-Curtis similarity is suggested (though you can change this), Section 5. If the data's first (or only) factor has some repeated levels, then a 1-way ANOSIM test is offered on that factor (Section 9). Standard group average CLUSTER (Section 6) is always suggested, though can be deselected, but the SIMPROF test option is greyed out if the ANOSIM box is ticked, so both cannot be requested in the same run of the wizard. (This makes good sense of course – if there is a predefined structure which you are interested in enough to want to test, then that is the primary test to carry out). If the ANOSIM box is not checked, a SIMPROF test is the default. The MDS box is always checked by default – this will be *n*MDS with the usual default options of 50 restarts, 2-d and 3-d ordination plots and Shepard diagrams (Section 8). The final proffered option is a SIMPER analysis (yet to be met – see the end of this section), either on the ANOSIM factor if that has been selected, or on the groups created by SIMPROF. If neither is chosen, then SIMPER is greyed out.

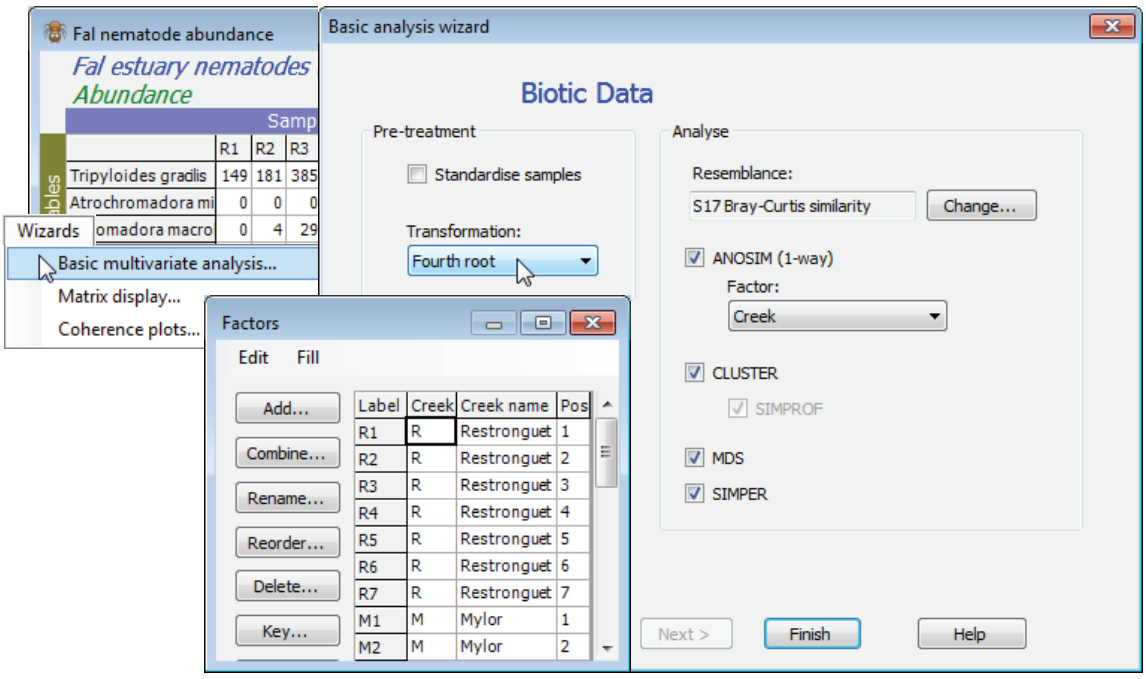
Basic MVA for structured data (Fal nematodes)

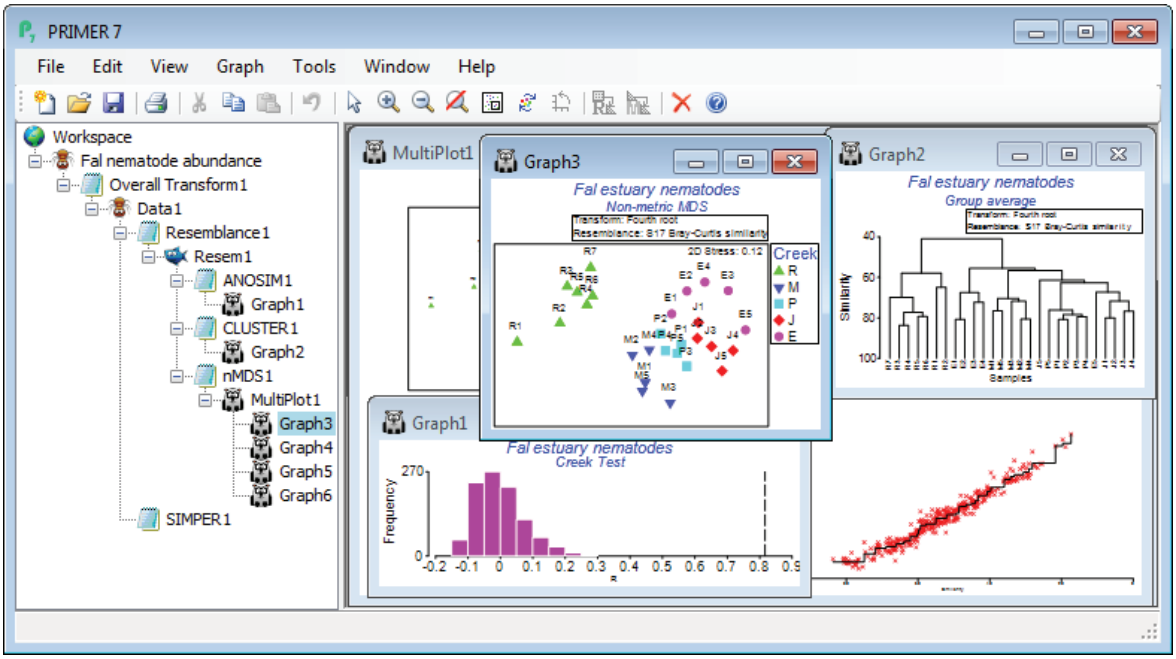
The benthic faunal study in the Fal estuary, Cornwall UK, was seen in Section 4. Sediment samples were taken at a total of 27 sites across 5 creeks running into the Fal estuary, with differing levels of heavy metal contamination from historic tin and copper mining – 7 sites in Restronguet (R) and 5 in each of Mylor (M), Pill (P), St Just (J) and Percuil (E) creeks. The existing workspace, Fal ws, concerns only the meiofaunal copepod community, and in Section 17 we shall see the macrofaunal component, but open here the nematode assemblages, Fal nematode abundance from C:\Examples

v7

v7\Fal benthic fauna, into a clear workspace (Fal ws2). **Wizards>Basic multivariate analysis** then offers the sequence of analyses shown below – in particular the routine picks up the existence of a *Creek* factor with repeated levels and checks the ANOSIM not SIMPROF boxes. In this case, a more severe transform is desirable to downweight the highly abundant *Metachromadora vivipara*, so change to (Transformation: Fourth root) – a shade plot (see later, and Section 4) helps here.

v7





v7 The main intention of the Wizards is to aid understanding of the steps involved, by examining the sequence of windows created in the Explorer tree. In this case, you can create precisely the same outcome by running the following. On Fal nematode abundance, take **Pre-treatment>Transform (overall)>**(Transformation: Fourth root) and, on the resulting transformed sheet Data1, **Analyse> Resemblance>**(Measure•Bray-Curtis similarity) & (Analyse between•Samples), giving Resem1. On Resem1, **Analyse>ANOSIM>**(Model:One-Way - A) & (Factors A: Creek)>(Type: Unordered) leads to ANOSIM test results, testing for significance of differences among creeks overall and pair-wise, in ANOSIM1, and a histogram of the null distribution for the overall (global) test, Graph1. On Resem1, **Analyse>Cluster>CLUSTER>**(Cluster mode•Group Average)&(✓Plot dendrogram) but without the SIMPROF box ticked, displays a standard UPGMA clustering in Graph2. Again on Resem1, **Analyse>MDS>Non-metric MDS (nMDS)**, with the defaults of (Min. Dimension: 2) & (Max. Dimension: 3)&(Number of restarts: 50)&(Minimum stress: 0.01)&(Kruskal fit scheme•1)& (✓Configuration plot) & (✓Shepard diagrams), gives the results window for MDS, nMDS1, which contains important information on how many times the lowest stress solution was observed in the 50 random restarts (if only a handful of times, consider running again with more restarts), and then

v7 the Multiplot1 window, which contains the (best) 2-d and 3-d ordination plots, along with Shepard diagrams. Clicking on any of these plots (or the + sign in front of the Multiplot1 icon) unrolls the four individual plots in the Explorer tree, Graph3 to Graph6. Finally, on the data sheet Data1, not the resemblance sheet Resem1 note, taking **Analyse>SIMPER>**(Design•One way) & (Factor A: Creek) & (Measure•Bray-Curtis similarity), with the other two boxes ticked, will give the detailed results window SIMPER1, breaking down the average dissimilarity between pairs of the ANOSIM groups (the creeks) into contributions from each species – see later this section.

The main conclusions here are that there are clearly significant differences overall between creeks (global R very large at 0.816, $p<0.1\%$) and, almost equally clearly, between all pairs of creeks – the only pairwise R value which drops as low as 0.5 is between St Just (J) and Percuil (E) ($p<2.4\%$). Other pairwise R values are in the range 0.75-0.99, and all of them are the largest obtainable values – greatest separation possible – in all permutations of the labels between the pair of creeks (i.e. 126 permutations for two creeks which both have 5 replicates, and 792 if Restronguet's 7 replicates are involved). In keeping with the pairwise ANOSIM R values for Restronguet (all $R>0.9$ even though the variability in Restronguet sites is large), the *n*MDS plot displays the very different nematode assemblages for that creek – it has by far the highest sediment concentrations of heavy metals.

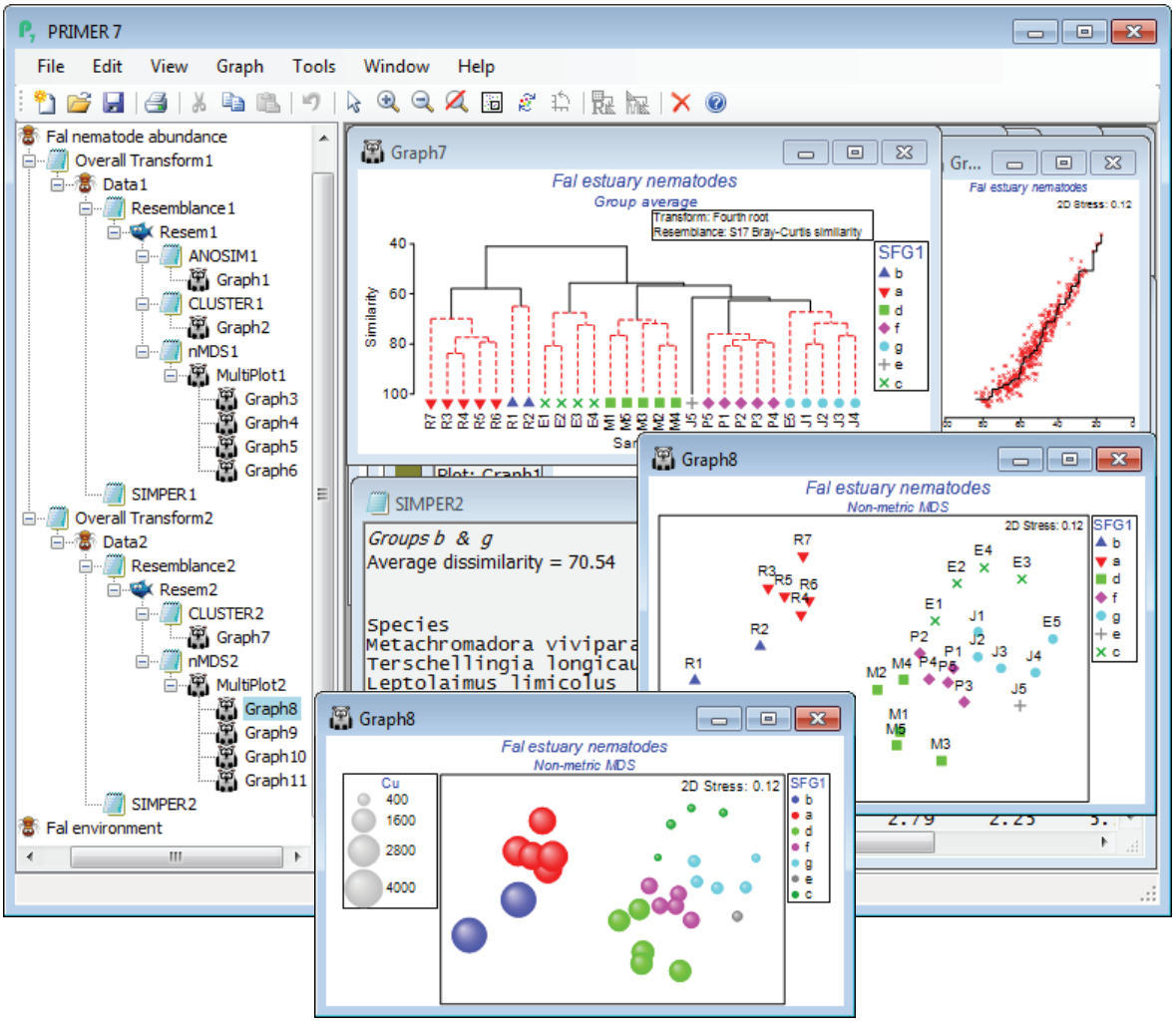
Basic MVA for *a priori* unstructured biotic data

In the Fal estuary study, where there are environmental data matching each of the 27 sites, it is not unreasonable to consider a different form of analysis, in which creek designations are considered secondary to the biotic relationships among sites in relation to the differences in the heavy metal concentrations (and such an analysis, by LINKTREE, is seen for this data in Chapter 11 of CiMC). In other words, we can choose to ignore the *Creek* factor and analyse the 27 sites as unstructured.

Another obvious example of this type of data, which we saw extensively in Section 6, is the zooplankton study of 57 sites in the Bristol Channel and Severn Estuary, which are laid out in a grid, and the primary analyses concern both a clustering of those sites into data-determined groups, and observation also (Section 8) of a more continuous gradation in the MDS plot – relating to salinity.

v7 Rather than re-open the Bristol Channel data again, we run through **Wizards>Basic multivariate analysis** once more on Fal nematode abundance but this time unchecking the (✓ANOSIM) box – the (✓SIMPROF) box is automatically now ticked. The output sequence in the Explorer tree looks similar (without the ANOSIM results of course) but now the dendrogram indicates the clusters of sites which are significantly separated by the series of SIMPROF tests (Section 6) – red dotted lines indicating sub-clustering which has no statistical support, with interpretable structure identified by black continuous lines. In fact, this largely accords with the creek designations – which was partly to be expected given the large pairwise R values in the previous ANOSIM test, but further division of the assemblages within creeks is more or less confined to two outlying sites in Restronguet. The SIMPROF test has automatically created a new factor, SFG1, of its 7 identified groups (one is the singleton J5) and this is available to all sheets and plots on this branch – and on the previous branch actually, since it back-propagates up to the Fal nematode abundance sheet and then forwards down the first branch – and SFG1 is displayed as symbols on the *n*MDS plot. In order to strengthen the visual association with the clustering, on the dendrogram (Graph7) you may wish to take **Graph> Sample Labels & Symbols**>Symbols(✓Plot)>(✓By factor SFG1). The SIMPER output is now in terms of these 7 SFG1 groups, rather than the previous Creek designations, naturally. It would also make good sense to superimpose individual heavy metal concentrations (or other abiotic variables) on the *n*MDS ordination, in a bubble plot. So, open Fal environment from the same directory, and with Graph8 as the active window, right-click over this (or **Graph**) and **Special>Main>**(Bubble ✓Bubble plot)>(Worksheet: Fal environment) & (Variables: Cu) – you will need to take **Change** to

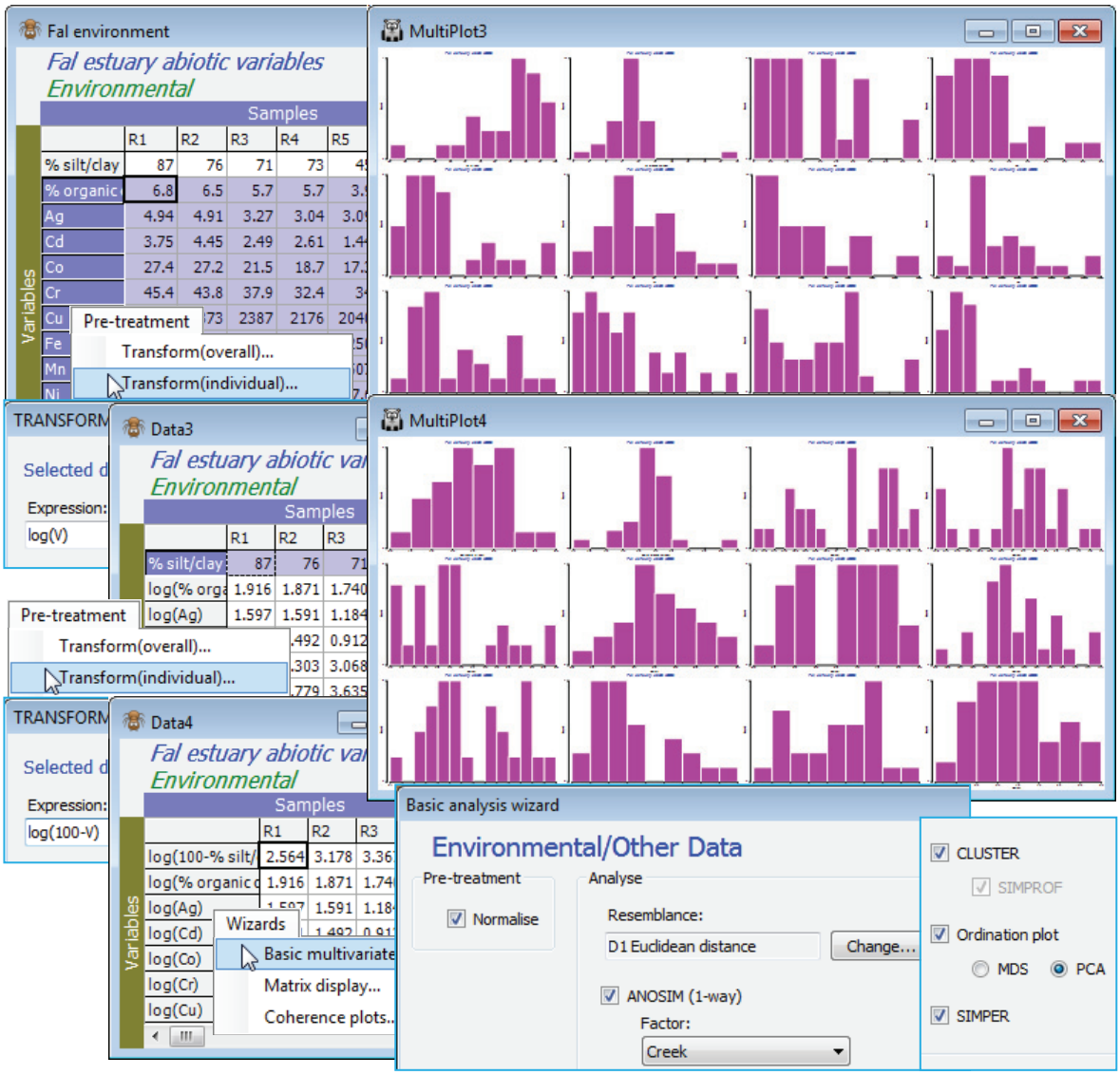
v7 select this variable, or you might prefer to Include multiple variables, for a segmented bubble plot. For a single variable, you might prefer the look of (✓3D effect) and, back on the **Graph>Sample Labels & Symbols** menu, uncheck the (Labels✓Plot) box and also (✓Plot history) on **General**.



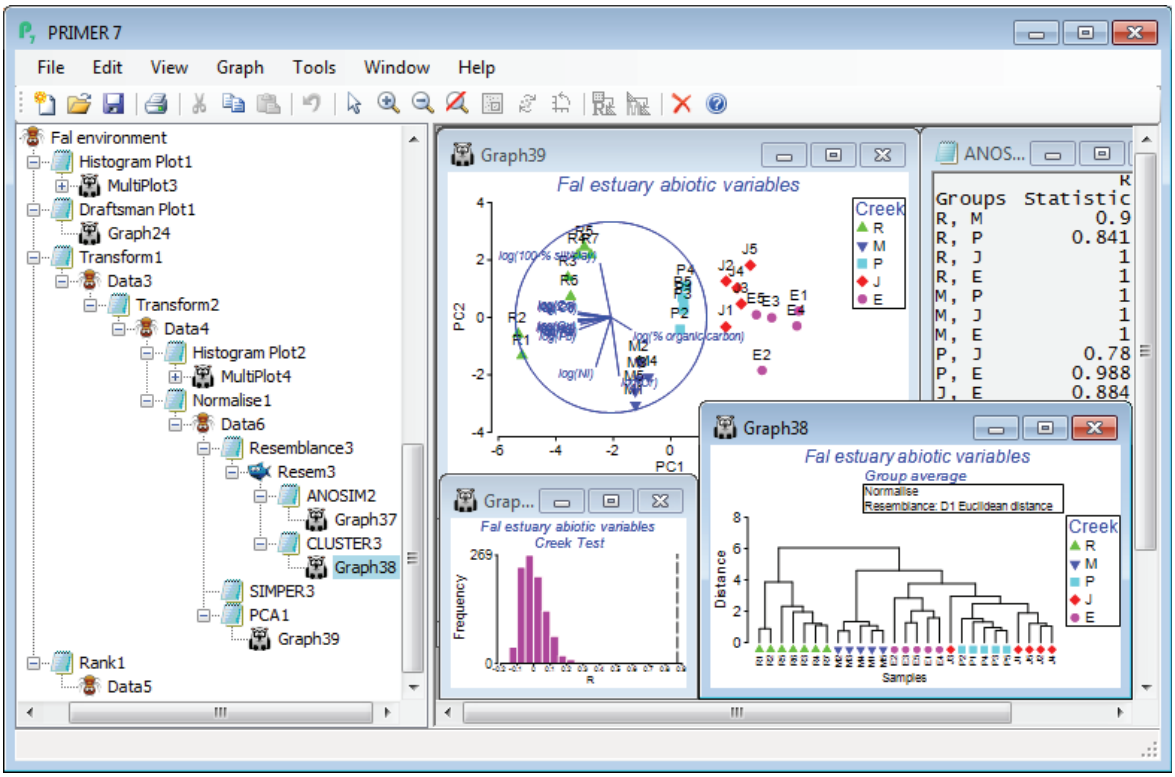
Basic MVA for environmental data

v7

Finally, try running **Analyse>Basic multivariate analysis** on this environmental data matrix, Fal environment, to look at the pattern in the abiotic variables collectively, rather than singly (a match of this multivariate environmental structure to the multivariate assemblage pattern is the basis of the BEST routine, Section 13 & 14). The environmental analysis it provides is fairly skeletal – the dialog box only offers one pre-treatment option, (✓Normalise), which would usually be taken since abiotic variables are typically on non-comparable measurement scales. However, here, as is often the case, the concentration variables would benefit from a transformation before getting to this stage – their distributions are typically right-skewed, as can be seen from **Plots>Histogram Plot** or **Draftsman Plot**. It would be optimal therefore to highlight all except the *%silt/clay* variable and take **Pre-treatment>Transform (individual)>**(Expression: log(V)), see Section 4, to give the new sheet Data3. The *%silt/clay* variable is of a very different type so it would not make sense to give it, automatically, the same transform as everything else. In fact, the histogram showed it to be left-skewed and Section 4 then suggests a transformation expression such as *log(100-V)*, or *log(101-V)* if the maximum value of 100 is attained for one of the samples. So, on Data3, highlight this first row and **Pre-treatment>Transform (individual)>**(Expression: log(100-V)), giving Data4. A re-run of **Plots>Histogram Plot** shows a set of transformed variables which are much less prone to the effects of outliers on the upcoming ordinations and tests, being fairly symmetric over their ranges. [If this pre-treatment stage seems all too much for you, at an early stage in your PRIMER experience(!), you could do worse than simply run **Tools>Rank variables** on Fal environment, which turns the 27 values for each variable into the ranks 1, 2, …, 27, and must totally remove the effects of any outliers, producing uniform distributions (at the price of loss of some sensitivity) – see under the Ranked variables heading in Section 11 and an example of the resulting draftsman plot in Section 12. This would be one of the (rare) occasions when the on entry to the Basic MVA routine, you do not take the default (✓Normalise) option, since all ranks are on the same scale.]



v7 Now, on the final, selectively transformed data sheet, e.g. Data4, take **Analyse>Basic multivariate analysis**, and because PRIMER has been told that this sheet is of Data type•Environmental (see the window's header line which will say *Environmental*, and if you need to change the type use **Edit> Properties**), the options offered by default will be (✓Normalise) & (Resemblance: D1 Euclidean distance), with the option to **Change** the latter to another resemblance measure. The analysis tools are now more or less the same as for biotic data, with (✓ANOSIM) proffered if a suitable factor exists (*Creek* in this case), and (✓Cluster), (✓Ordination plot) and (✓SIMPER). If ANOSIM is not checked, the default switches to (✓SIMPROF) tests on the standard clustering. The only difference now is that there is a choice of ordination options: (✓MDS) or (✓PCA), the former (again *n*MDS) being explicitly carried out using the supplied choice of distance coefficient, whilst PCA is only possible under an (implicit) Euclidean distance assumption. It follows that if a different distance measure has been selected, the PCA option is greyed out as unavailable. With ANOSIM run on the *Creek* factor and PCA for the ordination method, the Explorer tree under Fal environment is seen below. Note that PCA is run on the normalised data matrix, i.e. Data6 below, whereas for *n*MDS, the active sheet would have been the Euclidean distance matrix Resem3.



It is again instructive to repeat the same steps as **Analyse>Basic multivariate analysis** manually. On Data4, take **Pre-treatment>Normalise variables** (Section 4). The (✓Stats to worksheet) box is not ticked by default (if you check this, it just sends the mean and variance of each abiotic variable to a new sheet rather than listing them in the results window). On the normalised matrix, Data6, take **Analyse>Resemblance>**(Measure•Euclidean distance) & (Analyse between•Samples) to give Resem3, which is the active sheet for **Analyse>ANOSIM** and **Analyse>Cluster>CLUSTER**, both of which have exactly the same dialog as earlier, for analysing the biotic data in the *Creek* groups. As seen above, you may wish to add the *Creek* groups as symbols on the dendrogram with **Graph> Sample Labels & Symbols**. The other two routines start with the normalised data matrix Data6 as the active sheet. **Analyse>SIMPER** is set up as for the biotic data but with (Measure•Euclidean distance), which will be the default of course for data of environmental type. Finally, run Principal Components Analysis (Section 12) on Data6, with **Analyse>PCA>**(Maximum no of PCs: 5) and the other defaults – there is rarely any need to interpret more than the first 5 PCs. A vector plot (in blue) will automatically be overlaid – see Section 8 for the various vector plots available – but this can obscure the plot and is turned off, and on again, on the **Graph>Special>Overlays** tab with the check box (Vectors✓Overlay vectors)>(•Base variables).

v7 A run of Basic MVA with ANOSIM deselected again parallels the earlier options for biotic data.

The results show firstly that there are a lot of strong correlations among the abiotic variables, since the PCA results (PCA1) identify that the first 2 PCs account for 86.6% of the total variance and the first 3 PCs for 93.1% – these are very high figures. This is also seen in the eigenvectors, which give consistently large and negative values for all the metals (except *Cr* and *Ni*) on the PC1 axis, and negligible values on PC2. The vector plot shows these numbers graphically, with most metals thus increasing strongly towards Pill, Mylor and then, most strongly, the Restronguet creek samples (bubble plots would confirm this). In contrast, *%silt/clay, Cr, Ni* all have large eigenvectors on the PC2 axis, and relatively negligible ones on PC1, thus their vectors of increasing values point up or down the *y* axis (PC2) – *Cr* and *Ni* increase in the direction of the Mylor samples and Restronguet sites 1 and 2, as does *%silt/clay*. (Don't forget here that the silt/clay variable used was reversed to *100-%silt/clay* before taking logs, so *%silt/clay* increases down the page). The *%organic carbon* variable has its really large value on PC3, and this will largely account for the rise from 87% to 93% of the explained variation. Its contribution to the full multivariate abiotic pattern is not seen therefore on this 2-d PCA, though a rotatable 3-d PCA plot is simply obtained by **Graph>Special>** (Plot type•3D) and shows that site J3 largely accounts for this third axis.

The PCA also demonstrates clearly how the different creeks separate out in terms of their environmental variables, and ANOSIM formally confirms this, with a very large overall ANOSIM R of 0.87, reflecting very large pairwise R values also. This is getting close to the point (R=1, Section 9) where all Euclidean distances among samples in different creeks are larger than any within a creek. The cluster analysis is also seen to divide up by creek, more or less perfectly (again, excepting J3). This abiotic analysis therefore gives the basis for a correlative interpretation (likely to be causal, though not necessarily) of the similar patterns from the earlier run of **Basic multivariate analysis** on the nematode assemblage data – see Section 13 for more on linking biotic and abiotic analyses.

Wizard for Matrix display

v7

Another significant addition to the descriptive tools now available in PRIMER 7 is that of **Plots> Shade Plot**. This is particularly helpful in two ways. Firstly, we have already seen it used in simple form in Section 4 in relation to choice of transformation. A useful way of determining the effects of differing transformations – and thus which one might be adopted routinely for future analyses of particular community types in specific contexts – is to view differently pre-treated versions of the data matrix through shade plots (see Clarke KR, Tweedley JR, Valesini FJ 2014. 'Simple shade plots aid better long-term choices of data pre-treatment in multivariate assemblage studies'. *J Mar Biol Assoc UK* 94: 1-16). Shade plots are visual representations of the data matrix itself, in which the larger the entry in a specific cell, the darker the shade (or colour) plotted, white representing the absence of that species, and full black the largest entry (rounded up) in the whole matrix.

v7

Secondly, a shade plot can be an extremely useful tool at the later stages of an analysis, when the statistical tests have demonstrated the existence of structure, whether that is *a priori* sample groups or gradients tested with ANOSIM or RELATE, or whether a result of clustering and SIMPROF on unstructured samples, i.e. whenever we have licence from statistical testing to interpret the sample analyses in terms of individual taxa. If the rows and columns of the shade plot are re-ordered carefully enough, such a plot sometimes has a surprising amount of interpretative capacity, even if that is just to narrow down the set of species used in bubble plots on a sample *n*MDS. This is also the function of the **Analyse>SIMPER** routine (examples of which are given later in this section) but that has a focus on identifying species which contribute to differences among well-defined groups (as established by ANOSIM or SIMPROF), and is restricted to comparing pairs of groups at a time. It will therefore function poorly, if at all, for gradient structures where patterns of species change are more continuous – and where clear clusters of samples may not even exist. Shade plots can shed light on the reasons for both gradient and group structures and, for example, neatly distinguish between cases where a clear gradient on an *n*MDS plot is the result of a small number of species with strongly increasing or decreasing abundances across the full gradient, or a larger number of taxa which occupy quite different parts of the gradient, or whether in fact the multivariate summary comes from putting together information from a great number of species, each carrying limited or highly variable information – it is one of the triumphs of multivariate analysis that it is sometimes able to fashion a clear community structure from what would be most unpromising population data.

v7

The **Special>Reorder** button after **Plots>Shade Plot** offers a myriad combinations of clustering and re-ordering, both of rows and columns of the shade plot. The steps that are needed to exploit these to best advantage are quite complex, and **Wizards>Matrix Display** gets you started.

(Frierfjord macrofauna)

A data set not met elsewhere in this manual but which is used a great deal in CiMC, e.g. to start the description of 1-way ANOSIM permutation tests (Chapter 6), is from sub-tidal sampling by Day grabs for benthic macrofauna at 6 sites (labelled A-E, G) in Frierfjord/Langesundfjord, Norway. Four replicate grab samples were taken at each site, and the data matrix consists of counts of 110 macrobenthic species over the 24 samples (Gray JS *et al* 1988 *Mar Ecol Prog Ser* 46: 151-165). The file is Frierfjord macrofauna counts(.pri) in C:\Examples v7\Frierfjord macrofauna – though an Excel version of the same data sheet Frierfjord macrofauna (Excel)(.xlsx) is also in that directory, as a reminder of the format necessary to read Excel sheets into PRIMER (Section 1).

v7

So, save and close any open workspace (e.g. Fal ws2), open Frierfjord macrofauna counts and with this as active sheet, run **Wizards>Matrix display**, taking the options (✓Reduce species set>Keep most important: 50) & (Transformation: Square root) & (✓Retain sample groups>By Factor: Site). The outcome needs one minor change, but for the moment look at the display and note that:

a) the species have been clustered, to place together species which tend to have similar patterns of abundance across the samples, and the dendrogram rotated to attempt to diagonalise the matrix;

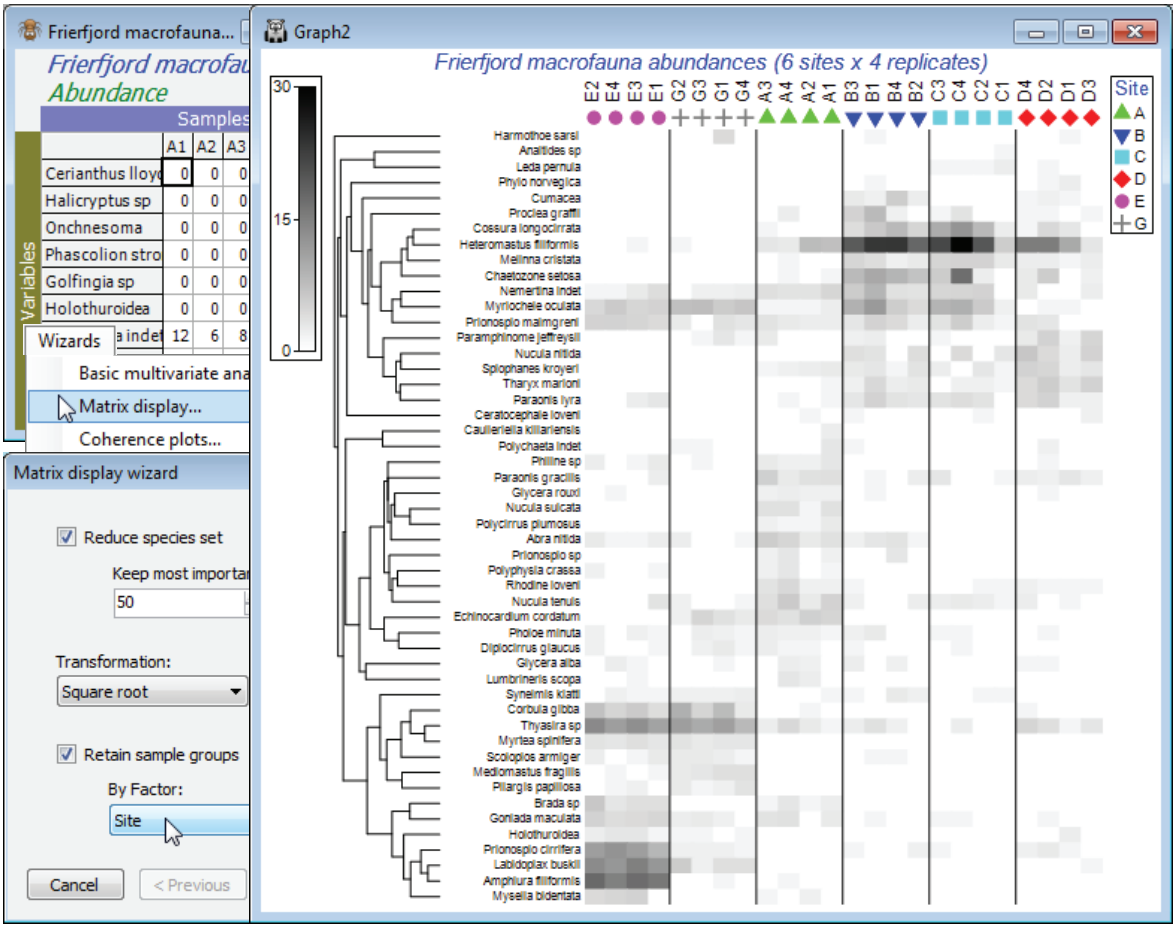
b) the samples have been ordered also to diagonalise the matrix, which is arguably unhelpful here, so a later tidying-up step will replace the site alphabetic order, to reflect the spatial layout; but

c) the replicates are kept together within sites by the (✓Retain sample groups) instruction and the vertical bars divide the levels of the specified factor (the sites);

d) the abundance of each species in each sample can be gauged from the depth of shading, but this is on the requested square root scale, so the (rounded up) maximum value on the scale bar of 30 corresponds to a count of 900, and 15 to a count of 225, the same scale bar applying to all species – the number of values and maximum of the scale can be altered by the user;

e) the resulting plot makes it clear why the MDS and pairwise ANOSIM tests (Fig. 6.2 in CiMC) split the sites into three groups: E & G (in Frierfjord above the sill, where there may be seasonal anoxia), B, C & D (in Langesundfjord below the sill, where any pollutants from industrial inputs at the head of the fjord system will be well mixed) and A (the reference site in Oslofjord); and finally

f) distinctions can be seen between E and G on the basis of the replicates and with more subtlety between D and the group (B,C) – ANOSIM (Section 9) does not separate these latter two sites.



v7 We now need to examine in more detail the decisions made and the routines employed to generate this Shade Plot, so that some limitations of the automated Matrix Display wizard can be relaxed. Note, however, that it is still often a good idea to produce an initial plot in this way, since many of the variations that are likely to be of interest can be achieved by one of the many options offered under the **Graph>Special** menu for shade plots, particularly the **Reorder** button.

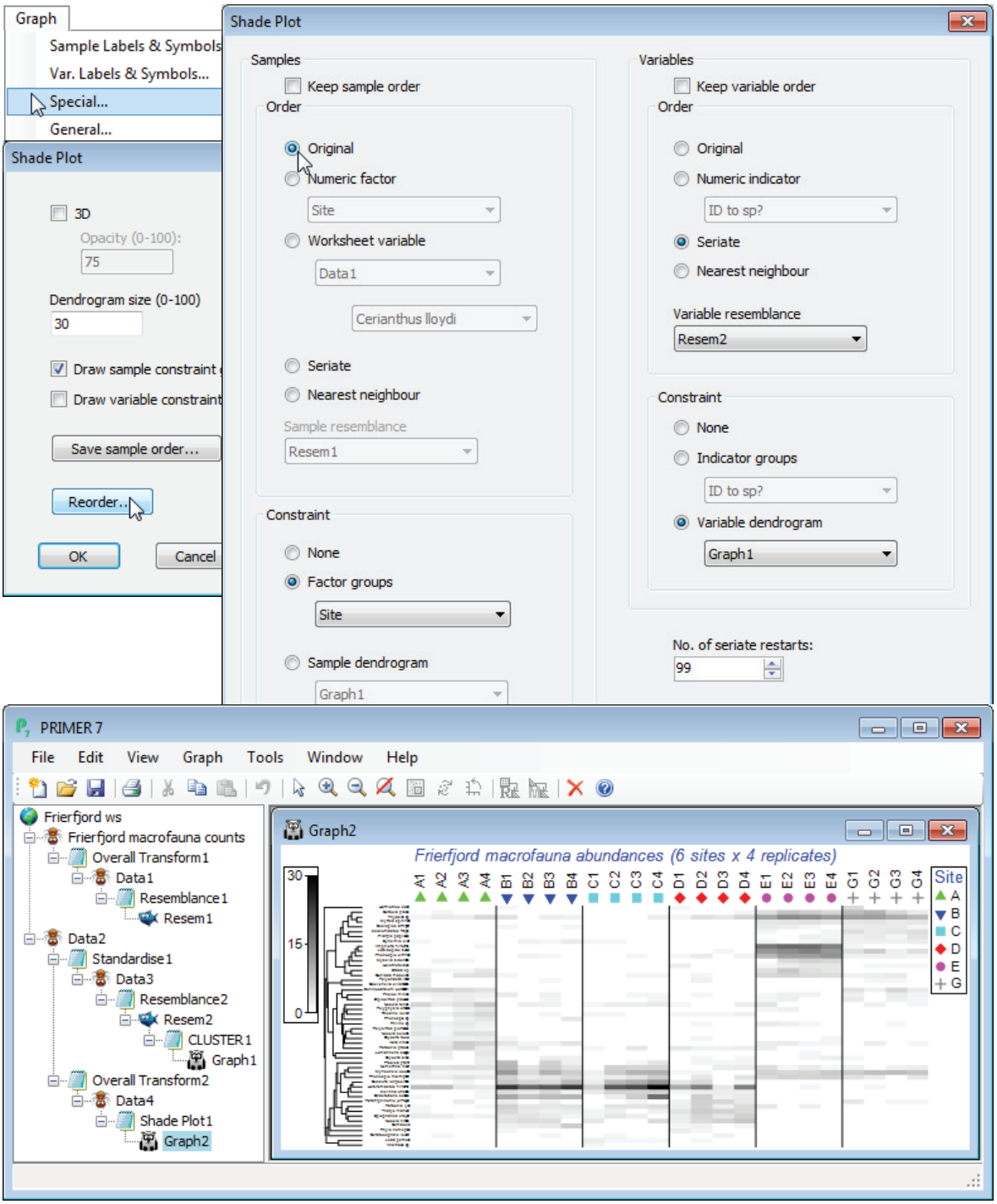
Reducing the species set

Of the 110 species, many occur only in one or two replicates, often as singleton individuals, so that whilst display of the whole matrix is perfectly possible, it will be cumbersome and less effective than viewing species which account for a non-negligible percent of the total number of individuals in each sample, and the Matrix display wizard dialog box default is (✓Reduce species set)>(Keep most important: 50). This concept was seen in Section 3 in the routine to **Select>Variables>**(•Use those that contribute at least 3 %), say. This would exclude any species which never (in any of the 24 samples) account for 3% or more of the total count for each sample. If that is run here it reduces the matrix to a set of 39 species. A weaker threshold criterion for elimination would be species not accounting for at least 1% of the total count somewhere, which leaves in 67 species. If phrased in terms of number of species retained, as in the option to **Select>Variables>**(•Use n-most important where n is 50) then the percentage threshold is manipulated until exactly 50 species are retained
v7 (this happens here if the % threshold is exactly 2%). This is the condition which **Matrix display** uses, and there is no flexibility to do other than change that threshold number of 50, but replicating the individual routines making up the wizard would allow a more flexible set of selection criteria, including Select>Variables>(•In at least n samples where n is). That you will need to do some species selection in large matrices is inevitable and often beneficial. It was stressed earlier that for sample analyses, all species can usually be retained (unless the resemblance measure involves a species standardisation, such as Gower or chi-squared distance) – the random nature of rare species occurrences in low numbers is given little weight in effective biological measures such as Bray-Curtis. But species analyses, defining similarities among species in their response over all samples – the idea of which was introduced in Section 5 – raise entirely different problems, and it is such
v7 species analyses that are the main topic of the remainder of this section. As seen above, in **Matrix display**, species similarities are used to cluster the species and/or re-order them in a way which optimises a seriation criterion and it is helpful if the rarer species, which cannot produce sensible assessments of similarity with other species – values will swing wildly between 0 and 100 – are deselected at the outset. Note, though, that it is an underlying principle of **Matrix display**, and thus preferably of direct **Shade Plot** runs, that where ordering/clustering of the samples is involved, it should be based on sample similarities from all species, not just those viewed in the shade plot.

Transforms in Matrix display

The next dialog box on **Matrix display** is (Transformation:), with default of Square root. This in effect runs the routine **Pre-treatment>Transformation(overall)** and thus offers the choices of None, Square root, Fourth root, log(X+1) and Presence/absence. The chosen transform is naturally applied both to the full data matrix submitted to **Matrix display** and to the reduced matrix (if a reduction is specified) which will form the entries of the shade plot itself. After transforming the full matrix, Bray-Curtis similarities will be computed among samples and used to determine the ordering and/or clustering of the samples on the *x* axis of the shade plot. On the *y* axis, the ordering and/or clustering of the (reduced) species set does not use this transform (or any transformation).
v7 The definition of similarity among species uses the Index of Association (IA), as defined near the start of Section 5, and this incorporates a species standardisation, which is usually calculated on untransformed data. The requirement for computation of IA on the original, untransformed data means that **Wizards>Matrix display** should generally be run on the raw data, and if a transformed matrix is submitted to it, this will elicit a warning box. You can happily ignore that if you wish to compute the similarities among species by standardising the transformed data – and there will be occasions when this is a reasonable thing to try. It would usually follow, of course, that when submitting transformed data you do not ask **Matrix display** for a further transformation step!

v7 Note that there is no facility for changing the choice of sample similarity away from Bray-Curtis, or defining similarity among samples by anything other than the Index of Association, within the **Wizards>Matrix display** routine. That is not to say that other measures might not be sensible or desirable in some cases but you will then need to create the files to be input into a direct run of **Plots>Shade Plot**. You can see the steps created by the wizard in the Explorer tree below, along with the **Graph>Special>Reorder** dialog needed to change the *x* axis to the original sample order.



Branches created in the Explorer tree

v7

The first branch takes the square root of the full matrix Frierfjord macrofauna counts, giving Data1, on which sample Bray-Curtis is calculated, Resem1. This was only used to seriate the *x* axis on the original shade plot but, as seen above, **Special>Reorder>**Samples>(Order•Original) in place of the default (Order•Seriate) has restored the axis to the label order of the data matrix. If the **Wizards> Matrix display** default of not retaining sample groups had been followed (no factor supplied), then Resem1 would be input to **Analyse>Cluster>CLUSTER**, creating a dendrogram (without running SIMPROF), displayed on the *x* axis and with Resem1 used to seriate samples within the constraints of dendrogram rotation. Resem1 is the right resemblance matrix to use for multivariate routines such as *n*MDS and ANOSIM. The second branch starts with a **Tools>Duplicate** copy (Data2) of Frierfjord macrofauna counts on which **Select>Variables>**(•Use n-most important where n is 50) has been run. It is species-standardised by **Pre-treatment>Standardise>**(Standardise•Variables) & (By•Total) to give Data3, on which **Analyse>Resemblance>**(Measure•Index of association) & (Analyse between•Variables) then gives the species similarities Resem2 on which CLUSTER is run in just the same way as it would be for samples. [The Standardise step is not really needed here

because IA will restandardise species again as part of its equation. It is included partly to remind you that there is a species standardisation step but also because there are other cases, such as the v7 Type 3 SIMPROF tests for *coherent species curves* (statistically distinguishable species clusters) later in this section, in which an initial species standardisation is required even though an index of association will be calculated afterwards, so this is a good habit to adopt. (The issue arises there because the permutation direction in Type 3 SIMPROF is across species, and this only makes sense if species are scaled to add to the same total).] The final sub-branch in the Explorer tree, off the data matrix Data2, with its reduced number of species, is the one that generates the Shade Plot. Data2 is transformed with the specified square root, to give Data4, which is input to **Plots>Shade Plot** to give Graph2. If you repeat that last step manually, you will see that the resulting graph is a simple snapshot of the data matrix with samples and species in exactly the same order as the input matrix and no clustering or other ordering of the axes.

Shade Plot options in Matrix display

v7 The Graph options taken automatically by **Matrix Display** to produce the final shade plot output are as follows. Firstly, **Graph>Sample Labels & Symbols>**(Symbols✓Plot)>(✓By factor: Site). Next **Special>Reorder>**Samples>(Order•Seriate>Sample resemblance Resem1) & (Constraint• Factor groups Site). The Seriate step – using the Bray-Curtis sample similarities from the full data Resem1 – was the one we switched to (Order•Original) above, at which point the Sample resemblance box is greyed out as not needed. The right of this **Reorder** dialog sets: Variables>(Order• Seriate>Variable resemblance Resem2) & (Constraint•Variable dendrogram Graph1), the latter thus displaying a dendrogram on the *y* axis which is the species clustering Graph1 we saw above. The seriate step also uses the species resemblances Resem2 on which this clustering is based, and the seriations on both axes (see below) need iterative processes from many different restarts, and this is set to (No. of seriate restarts: 99). It is important to realise, however, that when within the **Reorder** dialog any other worksheets or graphs could replace those computed automatically by **Matrix display**, since we are now within the **Shade Plot** routine, e.g. a different hierarchical clustering method could be shown (on either axis), such as the binary divisive **Analyse>Cluster> UNCTREE** – which of course you would have to run separately in advance of this **Reorder** stage, so that the relevant graphs are available. Another example of a separate analysis you might want to incorporate at this stage would be a species clustering which incorporated a (Type 3) SIMPROF grouping of species – see later in this section. Finally when you **OK** these steps and return to the initial **Special** dialog box, you will note that **Matrix Display** has ticked (✓Draw sample constraint group boundaries), which give the vertical divisions on the shade plot – these are determined by the Samples>(Constraint• Factor groups Site) step on the **Reorder** dialog.

You are likely to agree by now that **Wizards>Matrix Display** with its minimal three dialog boxes can save a lot of time and complexity compared with creating all the necessary sheets and graphs, and inputting those to **Plots>Shade Plot**! So it is usually worth using **Matrix Display** as a starting point and then amending the fine detail on the plot. More importantly, it ensures that robust options are taken so that you can interpret the Shade Plot with confidence that the data is being viewed in the form (in all essential respects) in which it is used by the multivariate ordinations and tests.

Seriate operation

v7 Slightly more detail on how the seriate options are constructed, on both samples and species axes, can be found in Chapter 7 of CiMC but will be described here also, since the concept of a *seriation model matrix* has not yet been met (see Section 14). There are two distinct, but related, Seriate operations. If a cluster analysis is not being displayed, the •Seriate option will attempt to place the objects (samples or species) in such an order, numbered 1, 2, 3, …, that when the distance matrix among those pairs of integers is calculated (the seriation model matrix), and that distance matrix correlated element-by-element to the resemblance matrix for those objects (using a non-parametric Spearman correlation coefficient ρ), the resulting ρ is maximised. A perfect seriation here (ρ=1) would be when the order of dissimilarities among (say) the species exactly matches the seriation model matrix – the further species are away from each other in the ordered list, the greater their species dissimilarity. In practice, of course, ρ=1 values will not be attained, but the idea is to get as v7 close to this situation as possible. However, to be sure that the optimum ρ has been found would require evaluation of all $p!$ arrangements of p species. Even when the species list is reduced to 50, as here, this is still 50! computations – an impossibly large number – so an approximate search routine is implemented starting from different initial random species orderings. Every time you enter the **Special>Reorder** dialog for a shade plot, and exit it with **OK** – whether or not you have

v7

made changes to the requested sheets or constraints – the number of random restarts specified, for either (or both) selected Seriate option(s), will be re-run. A different solution may then be found, so that the species or sample ordering changes slightly. It is worth experimenting with larger numbers of restarts to try to optimise the search, but in the end it is simply a slightly re-ordered display and a sub-optimal solution is likely to be just as useful for interpretation as a marginally 'better' one, so this is something not to be too concerned about. Note that you can switch the axes direction(s) by **Graph>Flip X** or **Flip Y** – these are arbitrary but in some contexts you may prefer a diagonalised shade plot, such as the initial one produced above, to run from top left to bottom right, rather than bottom left to top right. If you do not want to lose a reordering of species that seems to be visually helpful, when then going on to run **Reorder** options on the samples, you can fix the species order with (Variables✓Keep variable order) – or vice-versa, fixing the sample order in its current state by (Samples✓Keep sample order) before making changes to the species order.

Seriate a shade plot dendrogram

v7

The second Seriate option is when a dendrogram has been selected to be displayed on that axis – this is specified by **Special>Reorder>**(Constraint•Sample dendrogram) on the shade plot *x* axis or (Constraint•Variable dendrogram) on its *y* axis, supplying the appropriate dendrogram/tree plot graph window. Order•Seriate is still a possible option but it is *constrained* to be consistent with a dendrogram rotation. Section 6 described how the axis order for cluster dendrograms was arbitrary to within all possible rotations of the structure, viewed as a 'mobile'. But here the (Order•Seriate) option searches through those possible rotations for one which again maximises the correlation ρ of the resemblances to the seriation model matrix. This is a greatly reduced subset of the possible set of orderings in the unconstrained case – though still needing an iterative process – and is often more visually successful on a typically limited number of restarts. You can again **Flip X** or **Flip Y** on the resulting plot, but now also manually rotate the dendrogram after (or instead of) the Seriate option, exactly as you would for a standard cluster analysis, i.e. by clicking on the 'bars' of the mobile – these may be vertical or horizontal lines depending on whether the dendrogram is on the species or samples axis, respectively. After a careful manual rotation it may be particularly useful, again, to fix the current species ordering with (Variables✓Keep variable order) before tidying up the samples axis (or vice-versa). All such manual rotations are perfectly justifiable – in the end all we are doing is just looking at the data matrix! And a key point to remember is that multivariate analyses of samples (MDS, ANOSIM, RELATE etc) do not care about the species order in the matrix – they will return the same results, whatever the order. The human eye, however, does find it helpful to put together species with similar responses across samples – this can make the patterns that a multivariate analysis is able to pick out automatically suddenly become visually apparent!

v7

(Ekofisk oilfield macrofauna)

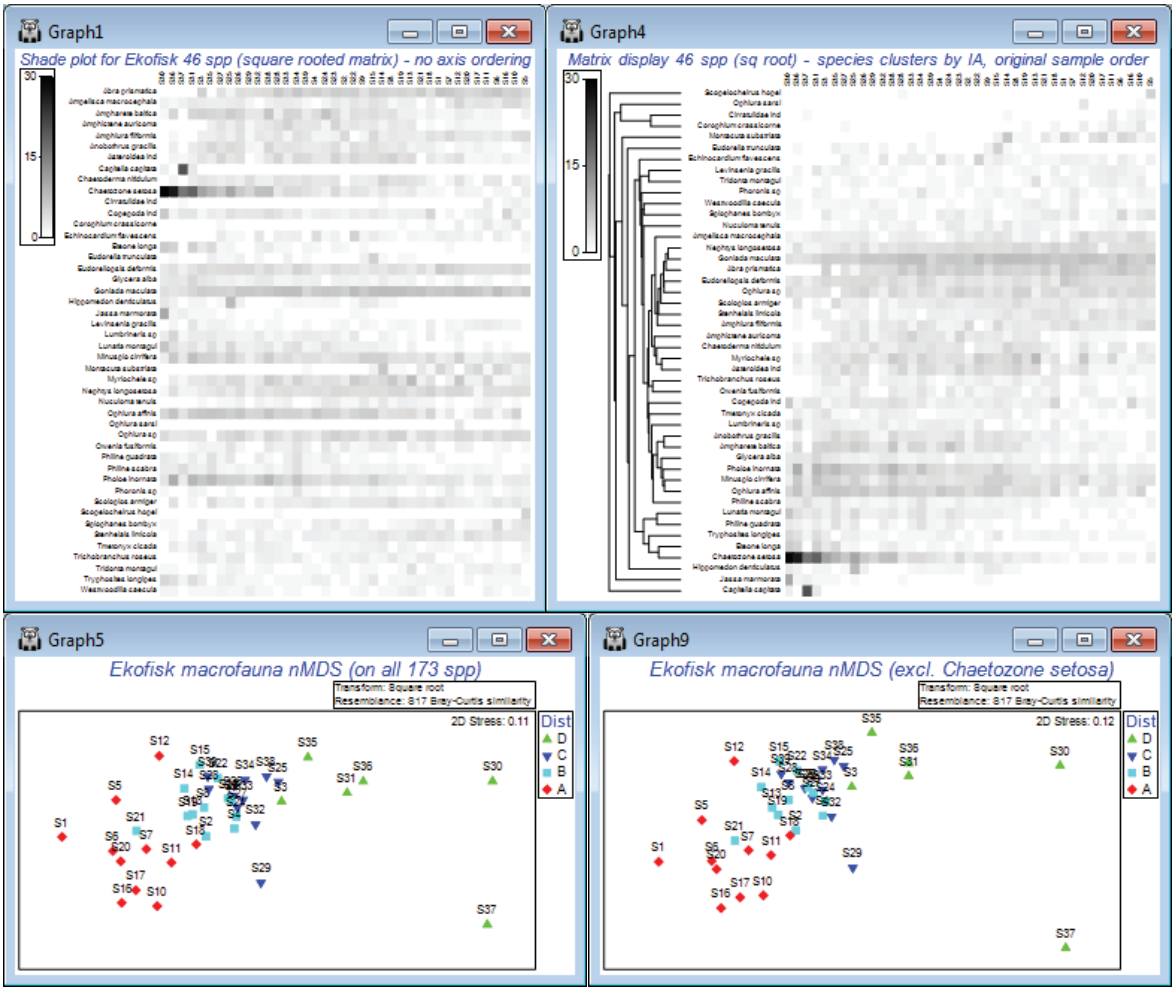
v7

A good example of how ordering of species in a shade plot can aid interpretation is seen for macro-benthic communities at different distances from the Ekofisk oilfield, last saved in Section 9 but which we have seen several times since it was introduced as the first example data set, in Section 1. These data have been run through **Plots>Shade Plot** before, in the discussion of transformation options in Section 4, where it was clear that unless some sort of transform was applied (square root looked the best bet), the effect of a single species (*Chaetozone setosa*) would dominate the result of a multivariate analysis. We remarked at the time that it was not immediately clear from the shade plot that, even after transformation, there would be a convincing relationship between the species communities and the distance from the oilfield of each of the 39 samples (note that the sample order in the matrix is of increasing distance from the oilfield, across the columns from left to right, and the species order in the rows is alphabetic, and this order will be unchanged by a simple run of **Plots>Shade Plot**). Yet we have since seen (in Section 8) a highly convincing *n*MDS plot of the gradient of change in community structure (on root transformed data and Bray-Curtis similarities) with distance to the oilfield, backed up by a strong 1-way ordered ANOSIM test result in Section 9. It might be argued from the shade plot shown earlier, and repeated below (left), that for the mild square-root transform, *Chaetozone setosa* will still have rather a strong input to this analysis – and it certainly has increasing values as sites gets closer to the oilfield. But the *n*MDS is repeated below after omitting this species and the convincing gradient pattern is little changed. Why this is so does become clearer, however, when we run the original data matrix through **Matrix display**, and re-instate the original sample order on the *x* axis – because of the clustering/seriation on the species axis we can now see groups of species which are: a) mainly restricted to the ends of the gradient; b) steadily increasing or decreasing with distance from the oilfield; c) in between those – species that increase then decrease along the gradient. It is a combination of all of these that gives a clear MDS.

v7

So, to avoid cluttering the existing Ekofisk ws for later (Section 14), open into a clear workspace Ekofisk macrofauna counts from C:\Examples v7\Ekofisk macrofauna and take **Select>Variables>** (•Use those that contribute at least 2 %) which retains 46 species, and **Pre-treatment>Transform (overall)>**(Transformation: Square root). On the result, run **Plots>Shade Plot** to obtain the left-hand plot below (as in Section 4). In contrast, **Select>All** on Ekofisk macrofauna counts, and take **Wizards>Matrix display>**(✓Reduce species set>Keep most important: 46) & (Transformation: Square root) and uncheck (✓Retain sample groups). The same 46 species are retained, as are the matrix entries, but species are re-ordered. So are the samples, which have now been clustered and seriated (subject to the constraints of their dendrogram) on the *x* axis, but this needs to be removed with **Graph>Special>Reorder>**Samples>(Order•Original) & (Constraint•None), whilst leaving the Variables side of this dialog unchanged. The only other change is to set (No. of seriate restarts: 999 or 9999). **OK** takes you back to the initial **Special** dialog – note that although (✓Draw sample constraint group boundaries) is ticked, no lines divide up the shade plot because (Constraint•Factor groups) was not specified on the Reorder dialog. The shade plot to the right, below, is the result. Under the shade plots are the *n*MDS ordinations, left: from all 173 species of Ekofisk macrofauna counts (root-transformed, then Bray-Curtis); and right: having omitted *Chaetozone setosa* by **Edit> Clear Highlight**, highlighting just that species, **Edit>Invert Highlight** and **Select>Highlighted**, then running root-transforms, Bray-Curtis similarities and *n*MDS in just the same way.

The interesting point about the re-ordered shade plot (right), which is worth re-iterating, is that the row and column orderings are independently derived. Looked at from the point of view of species rather than samples, the species index of association (IA) matrix is used to place species in their natural order, in terms of the differing similarities of their counts over samples, but it would be identical for any re-ordering of the samples. The sample ordering is fixed by an external factor – distance to the oilfield centre. Hence, if you can discern any diagonalisation of the matrix in this plot then that must be evidence for a serial relation of the community composition to distance from the oilfield. And that is what is now observed, and is picked up by the MDS and ordered ANOSIM. But, equally clearly, it is a subtle, whole community property that is being captured.



Close the workspace – save as Ekofisk ws2 if you wish, but we shall only need Ekofisk ws later.

(King Wrasse diets)

v7

A Western Australian study of the dietary assemblages of a single fish species (King Wrasse) were analysed as a 3-way crossed ANOSIM design, followed by an *n*MDS means plot, in Section 9. The three factors were samples taken at 3 *location*s (j1, j2, p2), at 2 *period*s in the year (S, W), and for wrasse of 4 *length*-class ranges (1-4), with 2 replicate (pools of) fish guts for each combination. Matrix entries were the percentage of the gut material by volume for each of 21 dietary categories, thus already sample-standardised, i.e. each sample adds to 100%. Square root transformation was taken prior to Bray-Curtis similarities, and the ANOSIM tests showed no effect of period at all (an average R value of 0.0), so the summary *n*MDS means plot averaged the (transformed) matrix over replicates and periods, to give 12 samples (3 locations by 4 length-classes). It is these averages that we will now input to Matrix display, to attempt to identify the dietary categories that it would be useful to display in a bubble plot on this averaged *n*MDS, which shows the rather weak location differences (average R = 0.26, p<1.5%) and stronger, ordered length-class effect (average R^O of 0.49, p<0.01%). In fact, this was the analysis that suggested a bubble plot of the *large crust*(acean) category seen in Section 9 – as noted earlier, other techniques for identifying contributing species, such as SIMPER (see end of this section), are not well suited to dissecting ordered changes.

v7

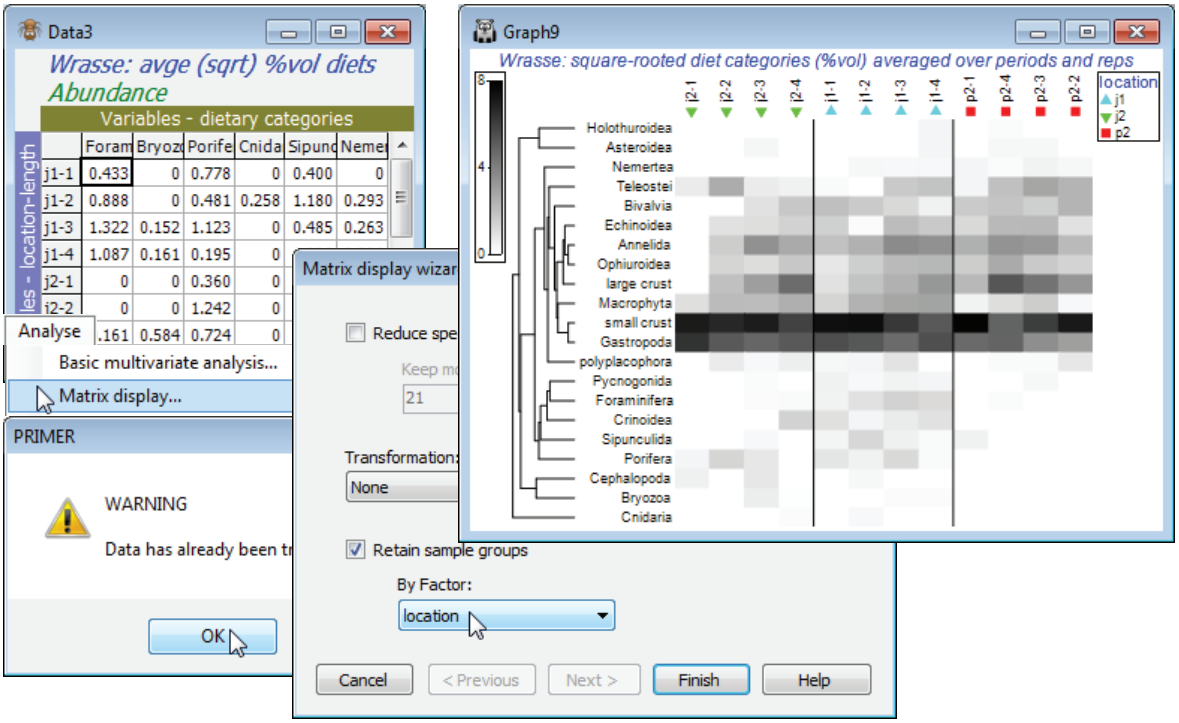
So, open Wrasse ws, or if unavailable, open Wrasse gut composition in C:\Examples v7\Wrasse diets and square-root transform this matrix. [As an aside, from the **Edit>Factors** sheet you will see how a **Combine** of the factors: location, -, period, /, length and rep has produced a combined factor then **Renamed** label, which was highlighted and copied to the clipboard (Ctrl-C), then the factor sheet saved, **Edit>Labels>Samples** taken, the existing labels (which were previously just integers 1, 2, 3, …) highlighted, and the more meaningful labels pasted over them (Ctrl-V).] We shall need a simpler **Combine** here under **Edit>Factors**, of location, - and length. Then run **Tools>Average>** (Averages for factor: location-length) on the root-transformed form of the matrix and enter this to

v7

Wizards>Matrix display. You will get the warning message: *Data has already been transformed* but this can safely be ignored for averaged data of this sort, where it makes good sense to do the transformation before the averaging (the issue of how best to create averages is briefly mentioned in Section 9 and again at the start of Section 17, but options are discussed in more detail in CiMC).

v7

The warning message is here to guide users towards running **Matrix display** on the raw data, and taking any necessary transformation within the routine, since it is (arguably) the preferred option for the species similarities to be computed on untransformed data, but this is by no means a 'hard and fast' rule, and here it is natural to run Matrix display on the (transformed then) averaged matrix – and not to request a further transformation, of course. Take **OK** on the warning message, uncheck (✓Reduce species set) since there are only 21 'species' (dietary categories), take (Transformation: None) and (✓Retain sample groups>By factor: location), to give the initial shade plot shown.

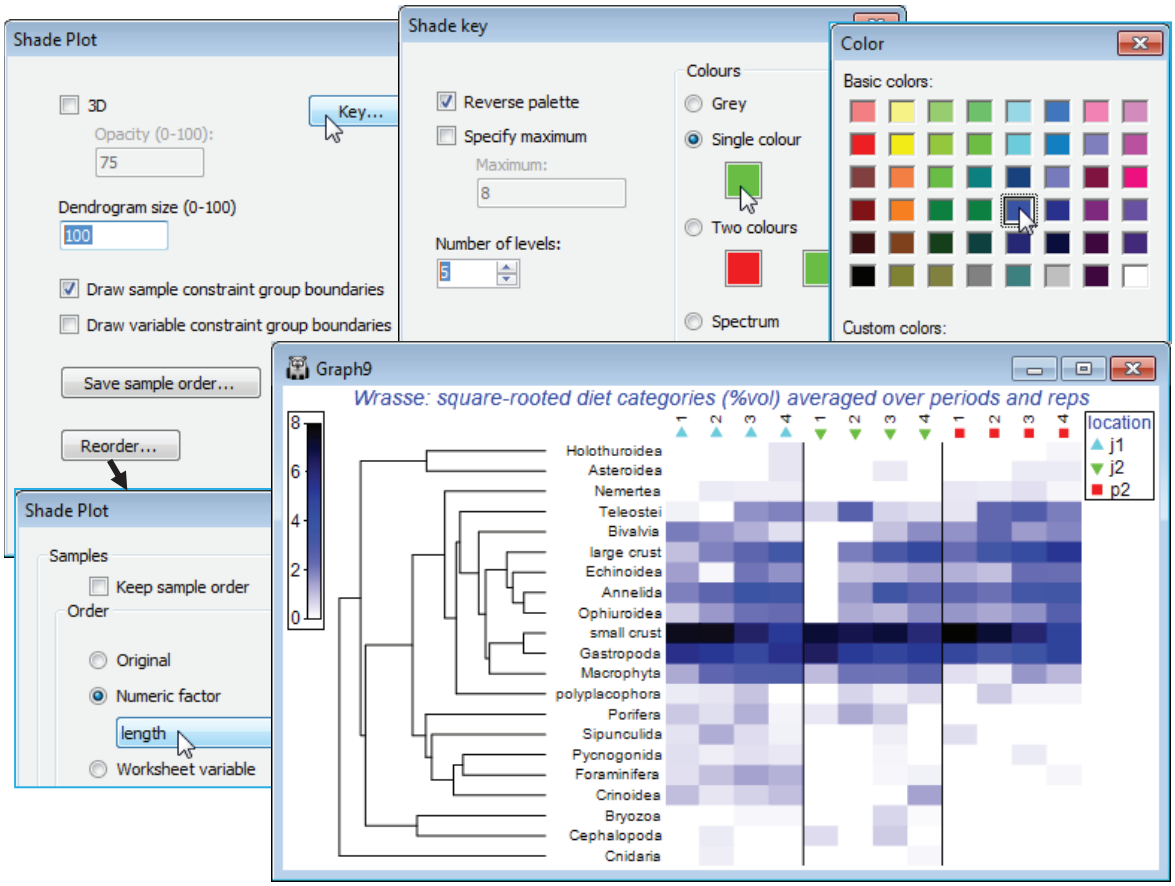


v7 Note that the submitted matrix had samples (location-length combinations) as rows, and variables (dietary categories) as columns, but the shade plot will always transpose the matrix in that case, to give a shade plot with 'species' as the *y* axis and samples on the *x* axis – there is no choice here. The resulting plot does still need some fine tuning, as usual, under three **Graph** dialogs (right-click over the plot). Firstly, **Samp. Labels & Symbols**>(Labels✓Plot)>(✓By factor length) would make the length categories more prominent. More importantly, these are seen not all to be in the correct

v7 sequence – their order was determined by the initial run of **Matrix display**, with its inbuilt attempt to diagonalise the shade plot (subject to the constraint of keeping location levels together). If we want to order by the *length* factor, it has to be numeric (we have seen this before with factors) and it is already numeric in this case – the size groups 1 to 4. So, on the **Special>Reorder** dialog, take Samples>(Order•Numeric factor)>length, leaving all other conditions unchanged, except to change the number of restarts to 9999 – this runs very quickly with the small number of species.

Special menu for shade plot

v7 With **OK** to get back to the first dialog screen under **Special**, we now look at the options available there. (Dendrogram size 100) will increase the size of the variables dendrogram (and a samples dendrogram at the same time, if one has been selected) so that its structure can be better seen – whether this is beneficial depends on how large a block of matrix entries needs to be squeezed into a sensible display area, but it is very useful if you want to manually rotate the dendrogram(s). The **Key** button leads to the Shade key dialog, which can also be accessed directly by clicking on the scale bar to the extreme left of the shade plot. This allows adjustment of the scale representing depth of shading, the specified maximum always corresponding to jet black shading(/colouring). It could be argued here that it might be useful to reserve this for the potential upper limit of the %vol scale rather than the default of a (rounded up) observed maximum. So, take (✓Specify maximum> Maximum: 10), since on the square-root scale used, this would back transform to represent 100% of average gut content being from one dietary category only. Also add some intermediate scale levels with (Number of levels: 6). The (✓Reverse palette) box would rarely be changed since this produces a shade plot in which black represents absence and white the specified maximum value. Black seems a less natural way of viewing absence in a community matrix, though it is the natural choice in a *heat map*, running from extreme cold to white hot. [Reference to this palette as reversed by default really comes from the fact that RGB scales have 0 values for all three colours for black, up to their maximum (255) for all three, producing white]. Other colours than grey for a shade plot are possible and the below shows a single colour choice, switched from the default of green to blue.



Shade plot colours

v7

You should experiment with the colour options, though there is not unlimited flexibility here in constructing a scale. A shade plot should ideally show a progression of colour which allows a natural feel for which are the larger and which the smaller entries in the matrix. For the (Colours• Two colours) option, the square to the left should be for the higher values (it will merge into black) and to the right for the lower abundances (it will appear from the white space absences). There is also a fixed colour spectrum choice (below). The other change made above was that the font for both the colour scale and symbol key benefitted from a small increase (both are of few characters). You can change Key font properties (as for any plot) by the **Keys** tab accessed from the usual **Samp. Labels & Symbols** or **General** dialog. For the above, this used **Keys font>**(Size: 120).

The corresponding *n*MDS ordination to the above shade plot is shown below – note the capacity a multivariate approach has to pick out the borderline significant (p<1.5%) location differences, from what the shade plot again shows to be rather subtle changes. The more striking (and highly significant) dietary progression with increasing wrasse sizes, running roughly in parallel for the three locations in the *n*MDS, is also evident on the shade plot. It is seen in the increasing consumption of a range of dietary categories (e.g. large crustaceans, annelids, ophiuroids, teleosts) and a decline in the dominant category of small crustaceans. An effective presentation is then a bubble plot on the MDS, thus linking individual prey categories to the pattern for the dietary assemblage as a whole,

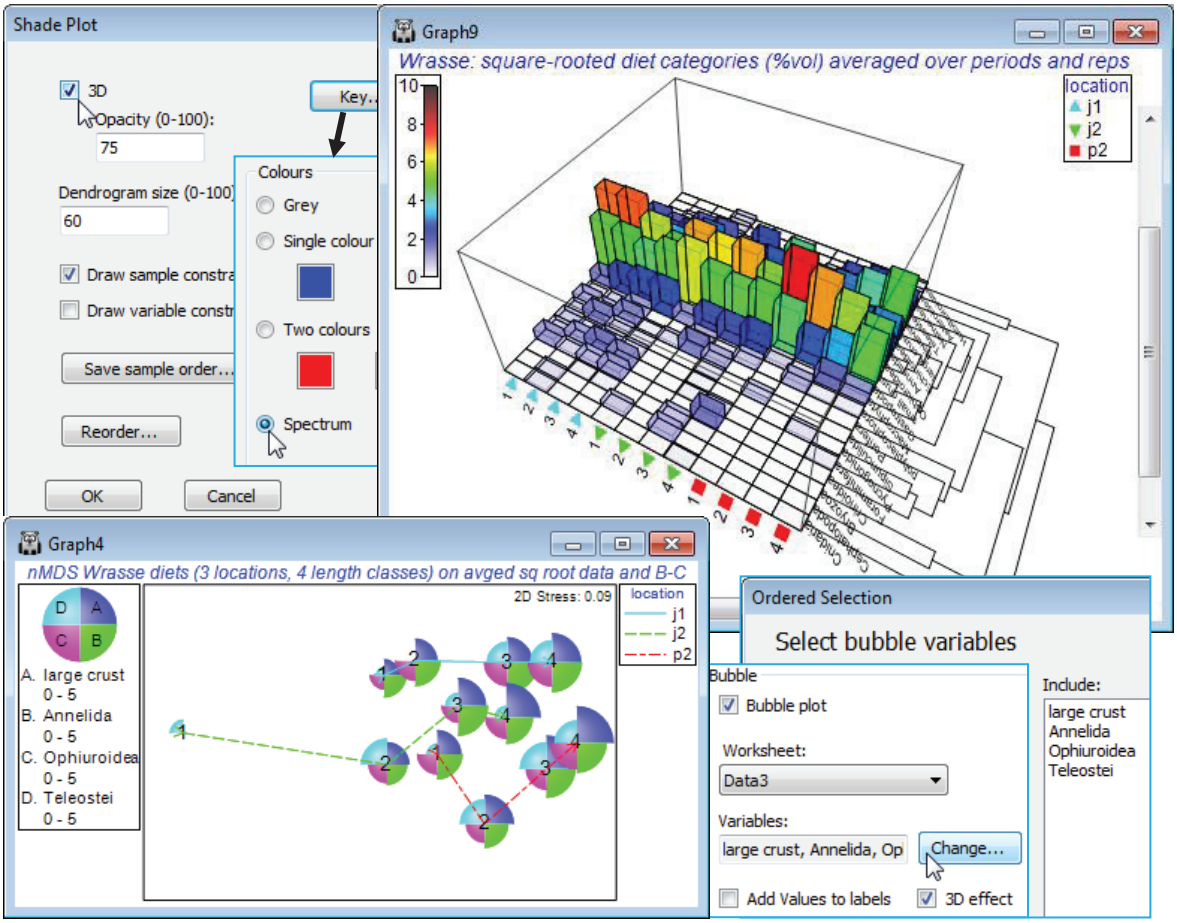
v7

and the ordination is shown below as a multiple bubble plot (Section 8). But the selection of dietary categories to display in this way (the ordered ANOSIM test having shown beyond doubt that there are such dietary differences to interpret) is greatly aided by careful study of the shade plot – which, importantly, does represent the matrix entries in the form they enter the similarity calculations.

3-d shade plot

v7

A further option on the initial **Special** menu for a shade plot (✓3D>Opacity (0-100): 75) turns the 2-d matrix into a 3-d bar plot, also carrying across any dendrograms and sample or variable labels and symbols on the axes. The (transformed) quantities in the cells of the matrix are represented in two ways, both by the height of the bars and by their colour, with the same scale key as for the 2D plot. The default value for opacity (75%) allows some viewing of bars through other bars, and smaller values, of course, lead to greater transparency (and an inevitably lightened colour scale). As for similar 3D plots, this can be manually rotated, or automatically spun with **Graph>Spin**, allowing the usual options for rotation speed and capturing the animation as an *.mp4 file etc.



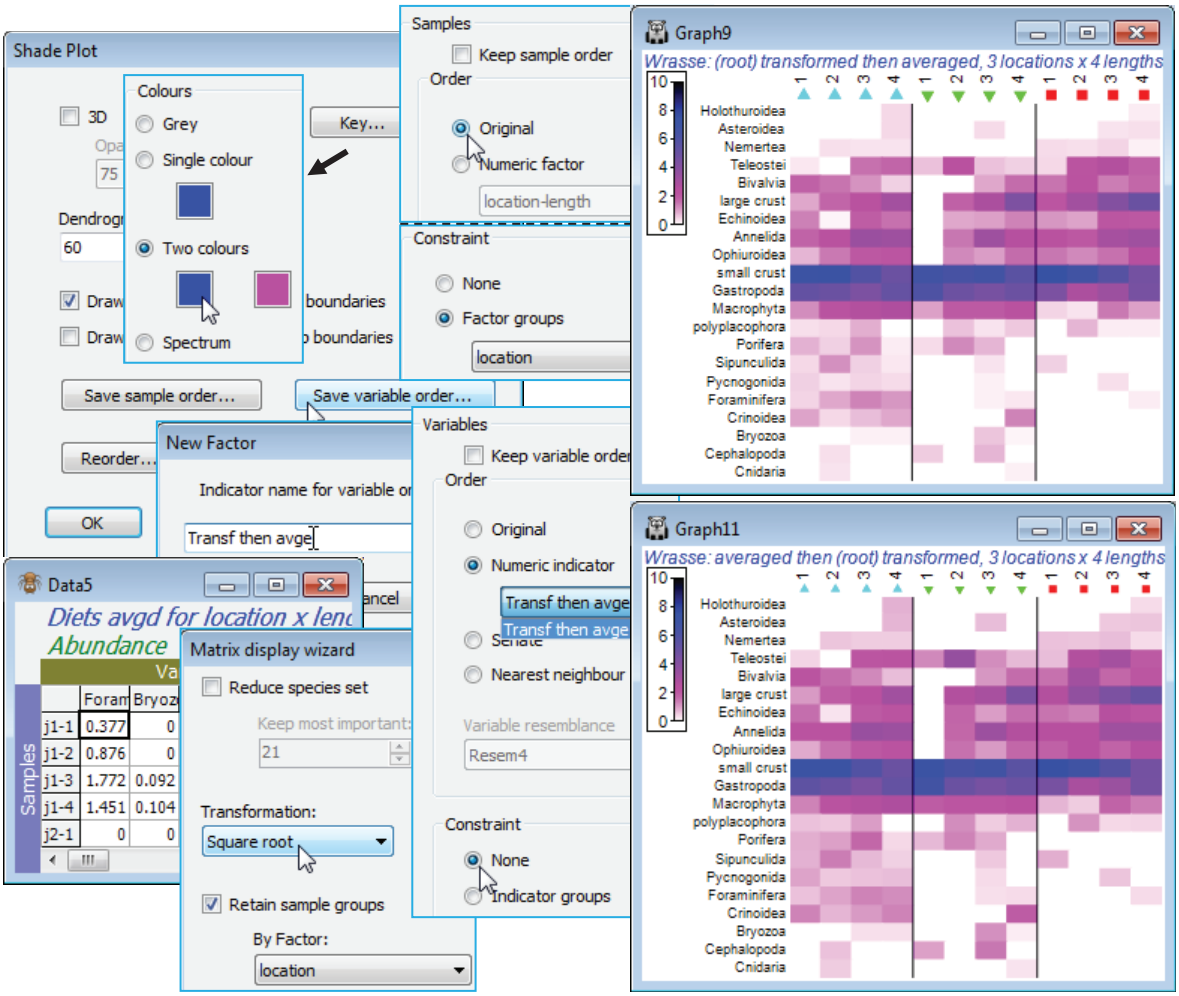
Save sample/ variable order

v7

A final choice on the **Special** dialog for shade plots can be useful when planning to compare plots. **Save sample order** and **Save variable order** allow the current order (top to bottom for species and left to right for samples) to be saved to a factor or indicator respectively. These could be used to re-order the original matrix (with **Edit>Sort>Rows** or **Columns**) for a direct run of **Plots>Shade Plot** – i.e. without the reordering which is implicit in **Matrix display** – or they could be used as an entry to (Order•Numeric factor) or (Order•Numeric indicator) on the **Reorder** dialog. An example of the latter would be to compare the shade plots produced by transforming then averaging (as performed above) with that obtained by first averaging, with **Tools>Average>**(Averages for factor: location-length) on the original Wrasse gut composition, then square root transforming the result in a run of

v7

Matrix display. Having run **Special>Save variable order>**(Indicator name for variable order: Transf then avge) on the first shade plot, under **Special>Reorder** on both plots take Variables> (Order•Numeric indicator Transf then avge) & (Constraint•None), and their species lists will align. The default plots from Matrix display are otherwise certain to have species in a different order, which will not make it easy to compare the two approaches. Note that a default run of **Wizards> Matrix display** under the (✓Retain sample groups>By factor: location) condition may also re-arrange the order in which the three locations are presented (since the attempted diagonalisation may differ). In general therefore, one could apply **Special>Save sample order** on the first array and **Edit>Factors>Import** this factor into the averaged matrix for the second array, thus ensuring that the two shade plots have both samples and species in the same order. In this straightforward case, however, it is simpler just to specify (Samples>Order•Original) under the **Special>Reorder** dialog for both plots, since we know that the matrix already has its samples listed in the required order (j1, j2, p2 locations, with length classes 1-4 within each, in that order). The comparison is given below, also demonstrating the other colour option not yet seen, mixing two colours between white and black. The conclusion is self-evident – the order in which averaging and a mild transformation (such as square root) are performed makes apparently negligible difference to the shade plot, and this is borne out in the multivariate ordination. In fact it is often observed to be the case that the various ways of displaying some form of meaned data do give very similar outcomes for balanced designs such as this – see the discussion in CiMC, Chapter 18. Save and close Wrasse ws.



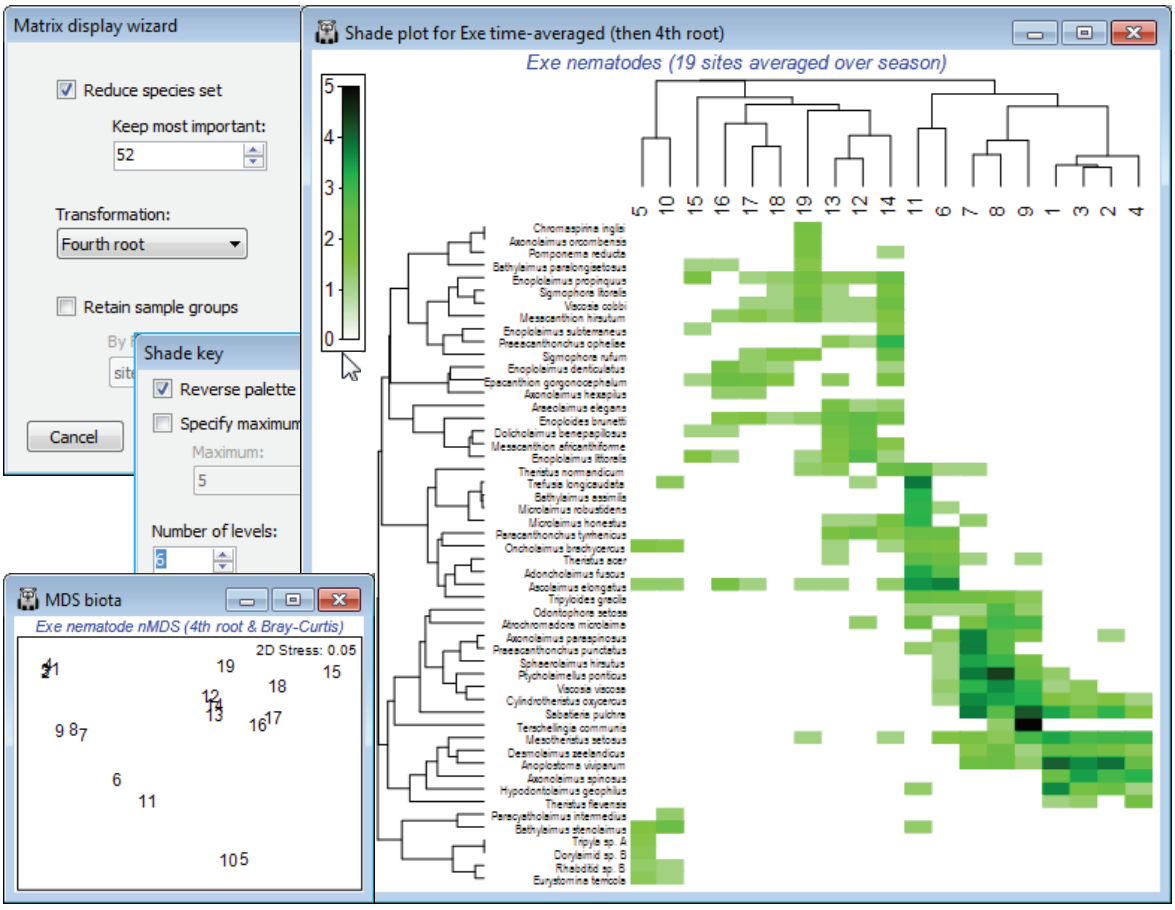
Clustering on species and samples (Exe nematodes)

v7

The time-averaged Exe nematode community data were used extensively in Section 8 to illustrate *n*MDS ordination, and the workspace Exe ws in C:\Examples v7\Exe nematodes should contain the data sheet Exe nematode abundance of the 19 sites, averaged over the 6 bi-monthly sampling times. This data matrix is now an example where there is no *a priori* group structure on the samples (sites) so that it would be instructive to cluster both species and samples axes of the shade plot. It is also a case of clear (and almost complete) turnover in species among some groups of sites, contrasting markedly in a shade plot with the more subtle community changes seen for the Ekofisk macrofauna and Wrasse diets. There are many species here which are infrequently found and in low abundance, and a useful shade plot would concentrate on only the 50-60 'most important', e.g. the analysis in Chapter 7, CiMC, picks only those species accounting for ≥ 5% of total abundance at any one of the 19 sites – using **Select>Variables>**(•Use those that contribute at least: 5 %) – which retains 52

v7

of them. So, on the active matrix Exe nematode abundance, **Wizards>Matrix display>**(✓Reduce species set>Keep most important: 52) & (Transformation: Fourth root) and uncheck the (✓Retain sample groups) box. The plot needs little tidying up: you might wish to put more levels into the scale key by clicking on the key and (Number of levels: 6), and in order to match Fig. 7.7 in CiMC, you may also need to **Flip X** or **Flip Y** (accessed by **Graph** or right-click). The dominant pattern is one of strong diagonalisation of the matrix but there is a point to note here. In the previous cases, the order of the sample axis was fixed on external information (a physical gradient and an *a priori* group structure, with ordered categories), and only the serial order of the species axis was given by optimising a matrix correlation ρ – but this seriation cannot be a function of the sample order since the species similarities would be unchanged by any sample re-ordering, hence observed diagonalisation implies a genuine gradient effect. When both axes are seriated as here (subject to constraints imposed by dendrogram rotation) then we need to be more careful in interpreting diagonalisation – it is inevitable that if both axes are internally ordered to an optimum degree, then the combination of them must appear diagonalised, at least to some extent. In fact, the *n*MDS configuration (Section 8) is not that of a simple linear gradient. Though the primary species in site group (5,10) could be rotated manually (by clicking on the dendrogram bar of the last species grouping) to sit at the top left rather than bottom left corner, this is arbitrary. The group is close to having an exclusive set of species – this would have made it 100% dissimilar to everything else and able to be placed at either end of the sequence (or even in the middle, if there were other sets of mutually exclusive species).



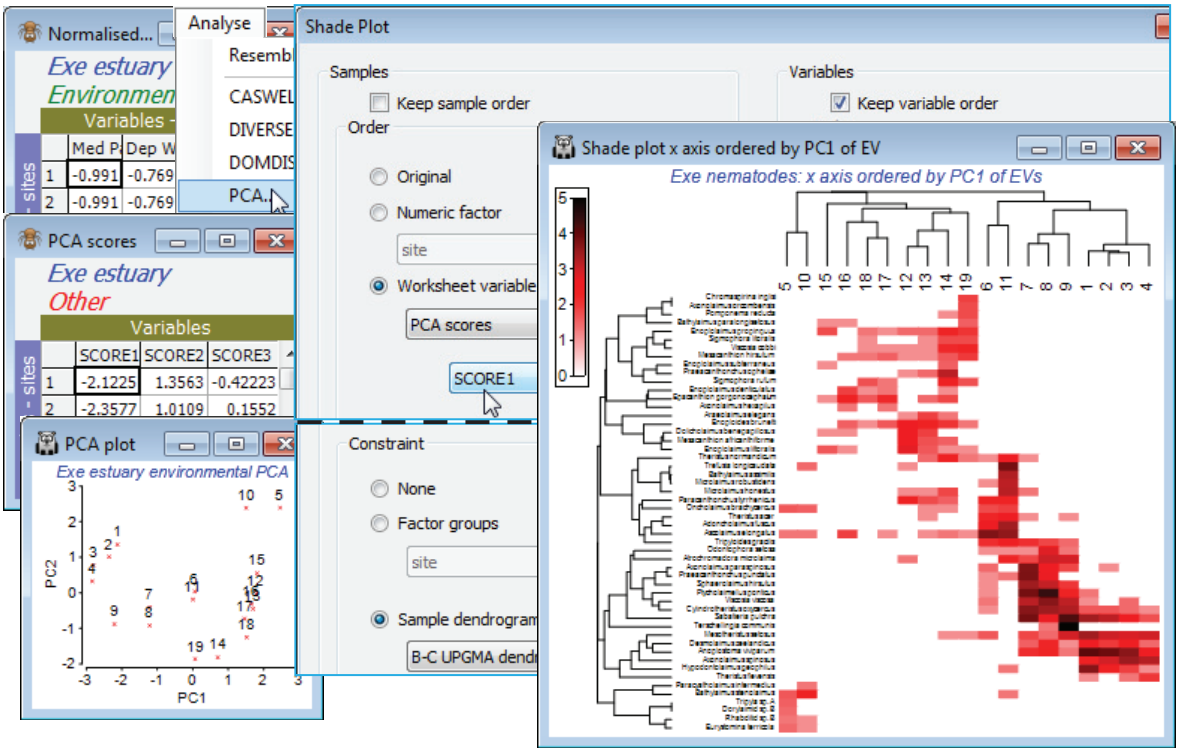
Ordering by a worksheet variable

v7

The core task faced by the **Special>Reorder** operation is to reduce what is inevitably a high-d similarity structure, on either axis, to a 1-d representation of those samples (or species). The default for **Matrix display**, when no group structure is provided for the samples, is to perform UPGMA (group average) clustering on the sample Bray-Curtis resemblance matrix, followed by optimising correlation ρ to a seriation model matrix. So the **Reorder** dialog for the above will have Samples> (Order•Seriate & Sample resemblance Resem*a*) & (Constraint•Sample dendrogram Graph*b*), with the same options for Variables (except based on group average clustering of Index of Association resemblances), but of course these are not the only possibilities. Any resemblance measure can be substituted on either axis, for the seriation, and any tree diagram for either dendrogram. We have seen the alternative of ordering by a numeric factor but (Order•Worksheet variable) also allows the sample axis to be ordered according to, for example, the value of an abiotic variable recorded over the same set of samples. This can be unconditionally or keeping samples from the same factor level together or only allowing rotations specified by a dendrogram – logically the dendrogram would be from the biotic matrix of the display. The latter is thus an interesting example of using both biotic and environmental data to produce a natural 1-d arrangement of samples not serially ordered by the biota. Apparent diagonalisation of the matrix can now be interpreted as a serial link between the selected environmental variable and the community pattern, without fear that we have 'chased the noise', since the sample clustering does not use the species order, or an internal seriation – instead it is the external abiotic variable that determines ordering (as opposed to grouping) of the samples.

As an example, open Exe environment if this is not already in the workspace – this manual has not made much use of these 6 abiotic variables, recorded for all 19 sites, though see Fig. 11.7 in CiMC. A logical single summary variable of the abiotic pattern might be the first principal component PC1 from a PCA ordination (Section 12). Take **Pre-treatment>Normalise Variables** on Exe environment – there is no need for an individual transformation of any variables here – and **Analyse>PCA** >(✓Scores to worksheet) on the normalised matrix. The PCA plot shows that, if we were to use the first PC (*SCORE1* in the sheet – renamed PCA scores – that this creates of the PCA co-ordinates), to order the samples in the shade plot unconditionally by this variable, we would not obtain a very meaningful sample order, either in terms of the (2-d) pattern of the biotic or abiotic samples – this is not a case of a single strong gradient (e.g. on PC1, site 10 would be split from 5 and placed in the middle of the loose 12-19 group). However, constraining by the existing biotic dendrogram (given by the original run of **Matrix display**) produces a neater, abiotic-driven dendrogram rotation, i.e. **Special>Reorder**>Samples>(Order•Worksheet variable PCA scores>SCORE1) & (Constraint• Sample dendrogram B-C UPGMA dendro) and (Variables✓Keep variable order). This last step greys out all the variable options and keeps the species list and dendrogram in the same order as above, avoiding a further 9999 seriation restarts – this can be useful in comparing sample options.

v7



Nearest neighbour ordering

v7

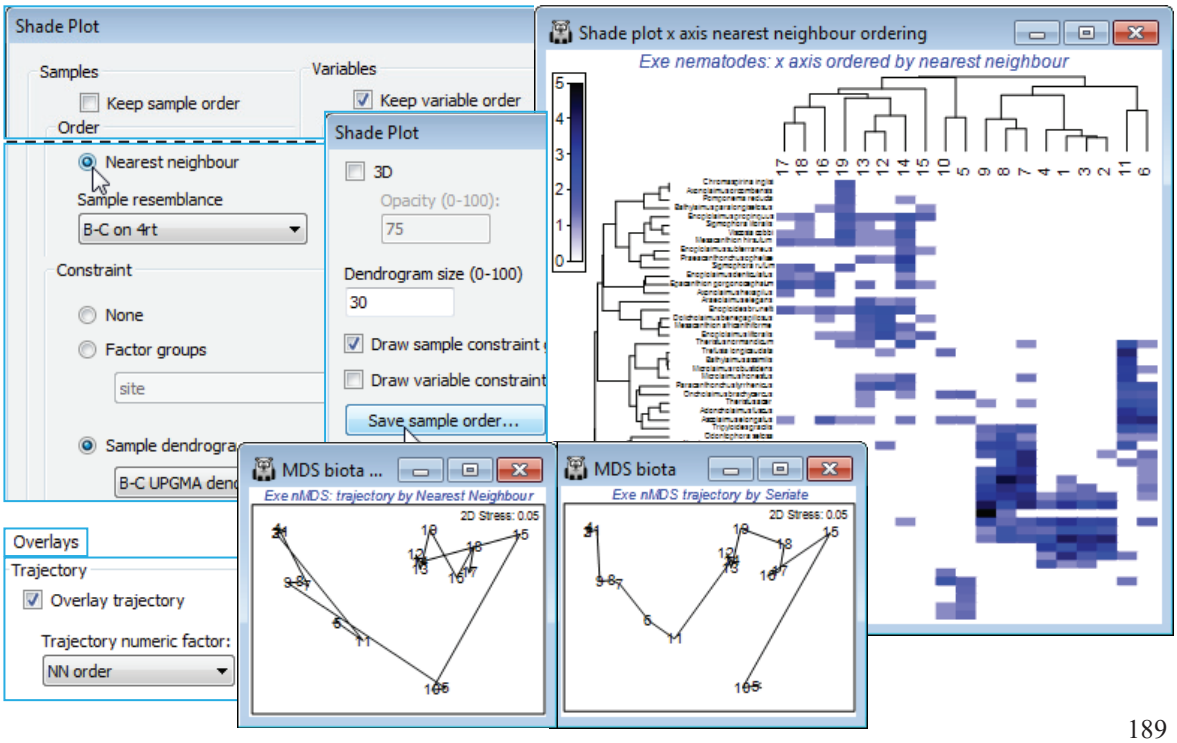
The final option that **Reorder** offers, attempting to find a useful ordering on a 1-d axis of a non-serial structure, is (Order•Nearest Neighbour) which implements a simplified *travelling salesman* algorithm. This is available for both samples and species axes, and an example where it is applied to both axes is seen in Fig. 7.9 of CiMC, but here for the Exe nematode data we shall run it on the sample axis only, again keeping the species order the same with (Variables✓Keep variable order). The full travelling salesman problem would be to find a route through the samples in which all are visited just once, and which minimises the total 'distance' travelled – in this context that means minimising the total dissimilarity between adjacent pairs of samples when placed in that order. We saw something similar with the minimum spanning tree in Section 8 (see box heading for MST), for precisely these Exe data. However, as the *n*MDS there shows, an MST allows branching of the route and this does not then provide a unique 1-d order in which to place the samples – but that plot does illustrate the point that a travelling salesman route should be greatly superior to, for example, taking the first axis of the 2-d MDS plot, when sites 5 and 10 would be interpolated in the 12-19 group (thus with unhelpful sample order in the shade plot of …, 12, 14, 13, 10, 19, 5, 16, …). The full travelling salesman problem is numerically highly demanding (a so-called *NP-hard* problem), even for modest numbers of samples, and PRIMER 7 currently implements the 'greedy' travelling salesman algorithm (see Chapter 7, CiMC), which is not an iterative process – so runs quickly – but may find an acceptable route, albeit possibly not an optimal one. It involves successive joining of nearest neighbours to one or other end-point of the route, hence run by (Order•Nearest Neighbour).

v7

For the existing shade plot of the Exe nematode data, **Tools>Duplicate** it (on the existing branch) and run **Special>Reorder**>Samples>(Order•Nearest neighbour>Sample resemblance B-C on 4rt) & (Constraint•Sample dendrogram B-C UPGMA dendro) and (Variables✓Keep variable order). The B-C on 4rt sheet is just the sample resemblances from the **Matrix display** run, used for all these shade plots – Bray-Curtis on 4th root transform of the nematode abundances, using the full set of species, which also gives the clustering B-C UPGMA dendro. You may need to **Flip X** to obtain the shade plot below. To note the order in which (•Nearest Neighbour) places the sites in relation to their (non-serial) pattern in the 2-d *n*MDS, on the shade plot take **Special>Save sample order>**(Factor name for sample order NN order), thus creating a factor which you can import (if necessary) into any sheet on the branch of the *n*MDS, with **Edit>Factors>Import**. On MDS biota, the 2-d *n*MDS, take **Graph>Special>Overlays>**(✓Overlay trajectory) and supply NN order. If you repeat this for the original serial order factor, the difference between ordering samples across the

v7

plot (•Seriate) or along a winding trajectory (•NN) is clear – and seen more starkly still in Fig. 7.9, CiMC – but the constraints imposed by the cluster analysis make either of these solutions better 1-d orderings for the purposes of shade plots than use of the first axis from a 2- or higher-d ordination. (Note that a 1-d MDS is not offered in PRIMER since each restart is sure to get trapped in a local minimum of the stress – there is limited ability to move points past each other in iterations in 1-d).



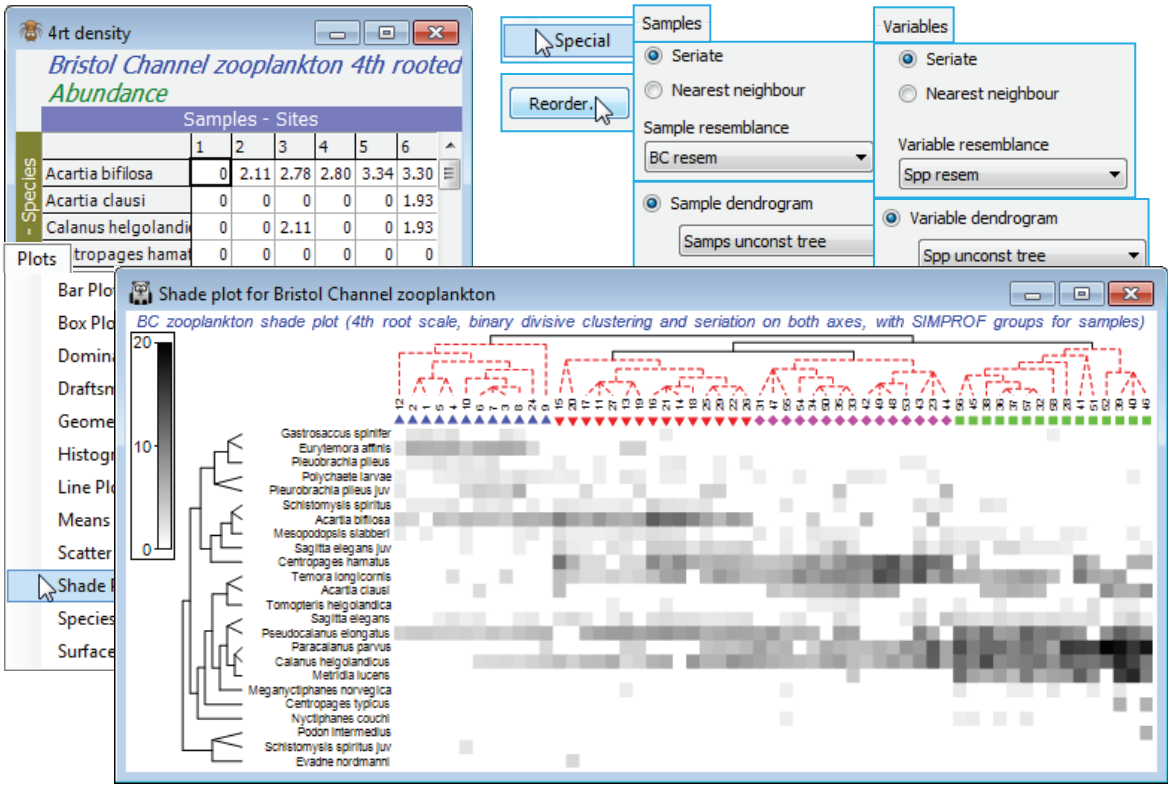
Other tree diagrams & SIMPROF (Bristol Ch. zooplankton)

v7

For the final example of the options in shade plot, save and close Exe ws and return to the Bristol Channel ws used extensively in Section 6 to illustrate different clustering methods (and seen again in Section 7 – and 8, for MDS bubble plots). If not available, open BC zooplankton density in C:\Examples v7\BC zooplankton, fourth-root transform it, calculate Bray-Curtis similarities BC resem and on this, run an unconstrained binary divisive clustering: **Analyse>Cluster>UNCTREE**>(Min group size: 1) & (Min split size: 4) & (Number of restarts: 50) & (Min split R: 0) & (✓SIMPROF test) & (Vertical positions•A%), and take defaults on the SIMPROF dialog, adding factor name Unctree which holds the SIMPROF group labels. Rename the dendrogram Samps unconst tree.

v7

Though it is usually easier to generate an initial shade plot from **Wizards>Matrix display**, it is instructive, for once, to create the components individually and input them to **Plots>Shade Plot**. The data matrix contains only 24 species and there is no real need to reduce it further (though three or four species are infrequently found, in low densities, so could be dropped without affecting the outcome). So, on BC zooplankton density (untransformed) take **Analyse>Resemblance**>(Measure •Index of association) & (Analyse between•Variables). This creates a species similarity matrix Spp resem which is input to **Analyse>CLUSTER>UNCTREE**, much as for the samples above except that the SIMPROF test is turned off this time (we shall see SIMPROF on species shortly), creating Spp unconst tree. With active matrix as the 4th-root transformed version of the plankton densities, 4rt density, run **Plots>Shade Plot** and on this **Graph>Special>Reorder**>Samples>(Order•Seriate >Sample resemblance BC resem) & (Constraint•Sample dendrogram Samps unconst tree) and then Variables>(Order•Seriate>Variable resemblance Spp resem) & (Constraint•Variable dendrogram Spp unconst tree), and (No. of seriate restarts 9999), the latter applying to the Seriate on both axes. Finally, add the sample SIMPROF groups as symbols with **Graph>Sample Labels & Symbols>** (Symbols>✓Plot>✓By factor Unctree) and you can remove (uncheck ✓Plot key on Unctree) or amend font sizes on the key from the **Keys** tab (**Keys font** and **Keys title font**). An alternative display would use, for the Samples, (Constraint•Factor groups Unctree) with everything else on the **Reorder** dialog unchanged, and (✓Draw sample constraint group boundaries) would then activate, to draw separating group lines, in place of the sample tree diagram. Other clustering methods such as the constrained binary divisive LINKTREE (Section 13) can also be used, and methods could be mixed on the two axes, if this seems desirable in a particular context. The shade plot here is clearly useful in identifying the key species that typify the four clusters of stations and which discriminate among them. This was used, along with the breakdown of species contributions of (dis)similarities among and within the groups (SIMPER, at the end of this section) to pick out key species for the (strongly serial) *n*MDS bubble plot of Section 8. Save and close Bristol Channel ws for use later.



Coherence plots wizard & Types 2/3 SIMPROF

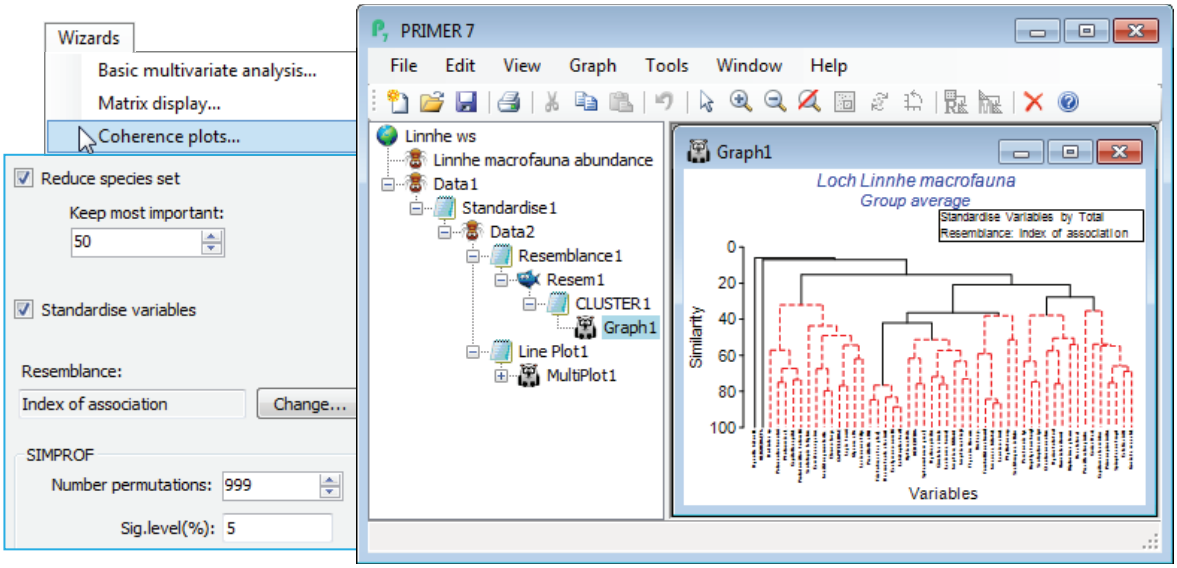
v7

The third item on the **Wizards** menu is **Coherence plots**. This is a combination of SIMPROF runs within a species **CLUSTER** analysis, to identify groups of *coherent species*, and **Plots>Line Plot**, to draw line graphs of species-standardised matrix entries (y) across samples (x), separately for each coherent species group. The output plots are therefore simple and transparent, describing patterns of response of each species across the sample ordering in the original matrix (e.g. through time, space or over an environmental gradient). The species abundance – or other quantity measure – is expressed in relative terms, as a percentage of total abundance for that species over all samples. The novelty here is that these line plots are grouped together in species sets which are statistically indistinguishable internally but significantly different over sets, using a series of SIMPROF tests, referred to as Type 3 SIMPROF. These are the precise analogue of similarity profile tests within sample cluster analyses that we saw applied in Section 6, to create significantly different sample groups (Type 1 SIMPROF) – and indeed one could trick earlier versions of PRIMER into carrying out Type 3 SIMPROF tests by defining species as samples and samples as species. This is now unnecessary (and was always rather confusing!) since PRIMER 7 automatically computes the right form of SIMPROF tests (1 or 3) in association with sample or species clustering. A single Type 2 SIMPROF test is also now offered in the **Analyse>SIMPROF** routine, which applies to species similarities (as with Type 3), but tests a different null hypothesis, namely that a set of species has no associations among any of its species (through competitive interaction or synergy, or more likely through opposite or common responses to differing abiotic conditions). In comparison, the null hypothesis for Type 3 SIMPROF tests is that the subset of species currently under test have a set of common (coherent) pairwise associations with each other. The technical distinction between the types is a combination of whether they calculate sample similarities (Types 1 and 4) or species associations (Types 2 and 3), and whether the tests work by randomly and independently permuting species over samples (Types 1 and 2) or permuting samples over species (Types 3 and 4). The Type 4 combination – testing sample similarity profiles by permuting samples over species is offered by **Analyse>SIMPROF** for completeness but would appear unlikely to have meaningful application. Some details on Type 2 and 3 SIMPROF tests on species are given in Chapter 7 of CiMC, e.g. the schematic diagram of Fig. 7.2, but the definitive paper on this, which also gives detailed discussion of several applications, is Somerfield PJ & Clarke KR 2013 *J Exp Mar Biol Ecol* 449: 261-273.

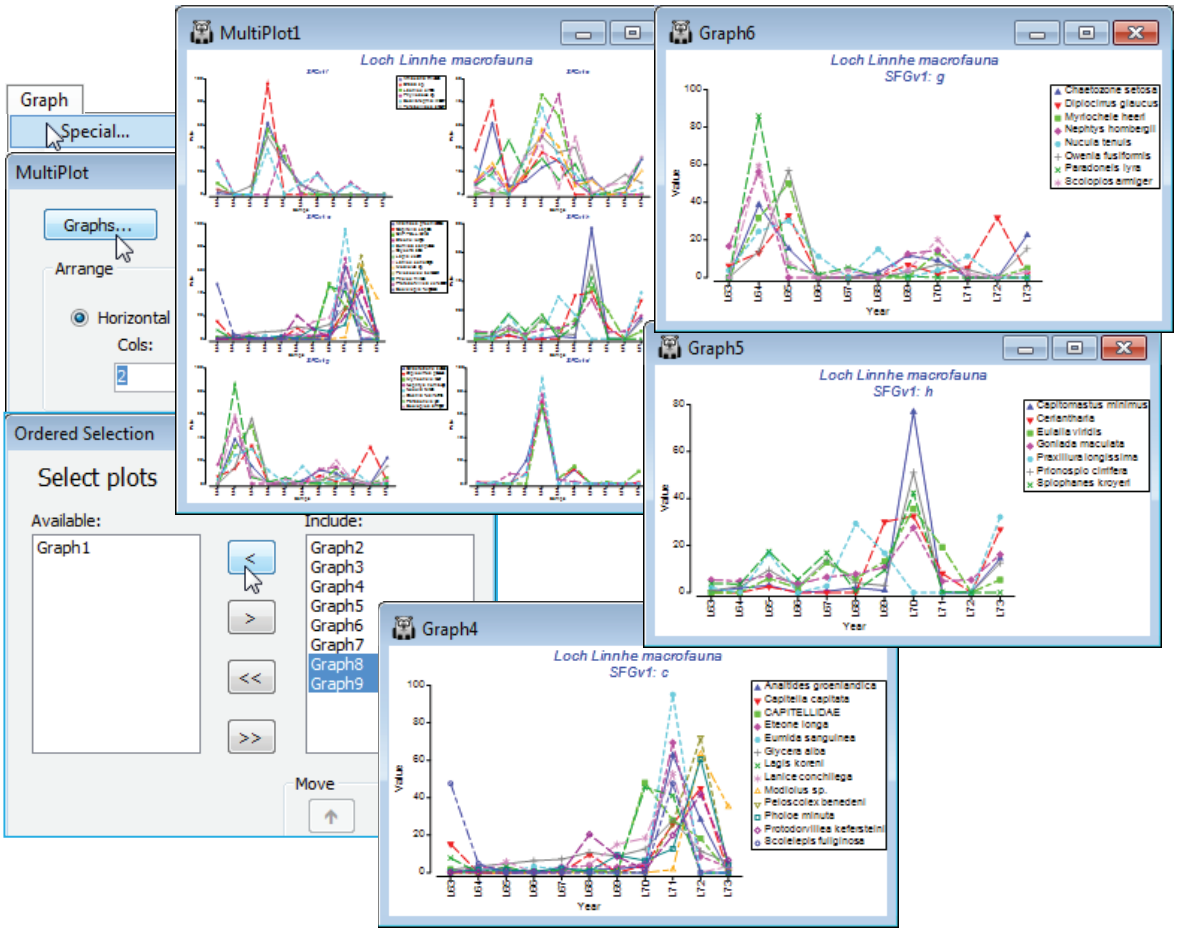
(L. Linnhe macrofauna)

v7

Soft sediment macrofaunal abundance at a single site in Loch Linnhe, Scotland was studied over 1963-73 by Pearson TH 1975, *J Exp Mar Biol Ecol* 20:1-41 (it is seen again, along with matching biomass data, in the species diversity curves of Section 16). The (pooled) data Linnhe macrofauna abundance, in C:\ Examples v7\Linnhe macrofauna, is of 11 samples (years) containing a total of 111 species. In 1966, pulp-mill effluent started to be discharged in the vicinity of the site, with the rate increasing in 1970 and reducing in 1972. On Linnhe macrofauna abundance, run **Wizards> Coherence plots**, taking the default settings shown below (50 species retains all those accounting for ~1% or more of the total abundance in at least one year, as seen from **Select>Variables**). The wizard first creates the selection in *Data1*, standardises the species, *Data2*, computes the index of association among species, *Resem1*, clusters this with Type 3 SIMPROF tests (creating group indicator *SFGv1*) and runs the line plots for those groups, held as separate graphs in *MultiPlot1*.



v7 The 'thumbnail' multi-plot is seen to contain 8 groups of line plots, of varying numbers of species in each, with two of the SIMPROF groups consisting only of single species which have a presence in only one of the 11 years (a different year). If you wish to remove these from the multiplot, as below, take **Graph**(/right-click)>**Special>Graphs** and send *Graph8* and *Graph9* to the Available rather than Include box. Also use this menu to change the layout to, for example, 3 rows by 2 cols, by (•Horizontal>Cols: 2). Clicking on an individual plot within the multi-plot unfurls the full set in the Explorer tree, and each plot identifies in its subtitle the letter (a, b, c, …) giving the level for that group in the indicator *SFGv1*. From the reduced set in Data1, the entries for just that group of species could be examined by selecting them with **Select>Variables>**(•Indicator levels)>(Indicator name: SFGv1)>**Levels**. Note that in Data1 this would be the original abundances – the standardised values used in the plot could be selected in the same way from Data2. [And if selection was needed from the original Linnhe macrofauna abundance sheet, you would first have to **Edit>Indicators> Import>**(Worksheet: Data1)>**Select**>(Include: SFGv1) from that original data matrix to import the SIMPROF groups, since the reduced matrix selection in Data1 is on a new branch of the Explorer tree and any factors and indicators are never automatically propagated between different branches].



The 6 major *coherent species sets* identified by the SIMPROF tests span a range of responses to the effects of the pulp-mill effluent from groups of species whose abundances: (*g*) are largely confined, in relative terms, to the earliest, pre-impact, years; (*f*) peak at the earliest stages of impact but then decline; (*d*) have a similar pattern but not kicking in until a year or so later; (*h*) show the same peak and decline but much further into the impact sequence; (*e*) stay relatively abundant until the most impacted years of 1970-72; and (*c*) are mainly the real opportunists (such as the Capitellids) that thrive in the most impacted years but decline sharply under the ameliorated conditions of 1973 – a year in which some of the other groups show signs of bouncing back. [Note that the graphs could have been arranged in the order described here, within MultiPlot1, again by **Special>Graphs** but now using the **Move** arrows on the Include list. You may find it helpful to rename the individual graphs *Graph2, …, Graph7* with their SIMPROF levels, respectively *f, e, c, h, g, d*, but note that you will then have a blank multi-plot since it can no longer find the original names(!) – this is fixed easily by **Special>Graphs**, highlighting f to d and moving them across together to the Include list].

v7

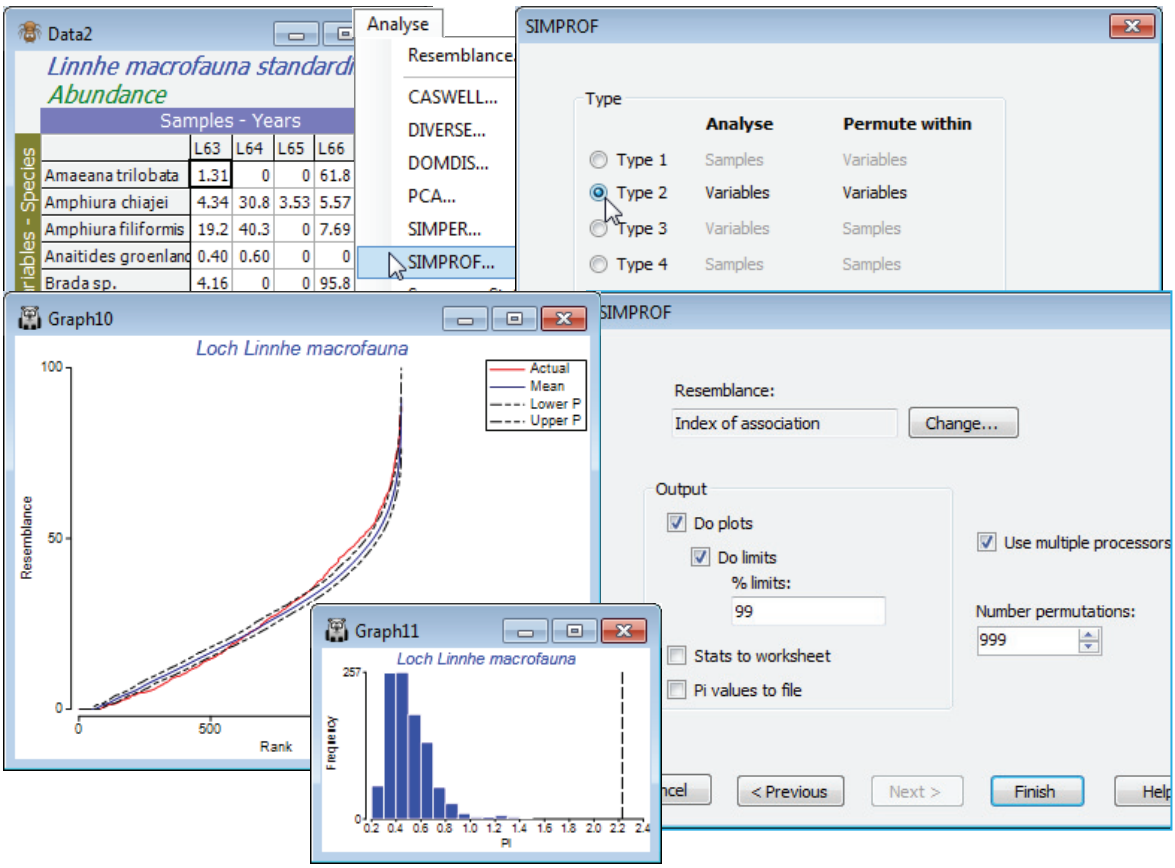
Running Type 2 SIMPROF

v7

A Type 2 SIMPROF test is not part of a **Wizards>Coherent plots** run, and there is a good case for carrying out a test of its null hypothesis (H_0: there are no species associations at all) prior to trying to break down those associations into coherent groups, i.e. within each group the null hypothesis (H_0: all species similarities within a group are the same) cannot be rejected. Inclusion of species with so little information that species similarities (index of association, IA) are totally unreliable is again unhelpful, so we start from the matrix reduced to the 50 'most important' species. For this test (Type 2), it does not make any difference whether we use the selection in Data1 or its species-standardised form Data2, because the permutations will be across samples within each species and IA includes a standardisation step in its formula. (It does, however, matter a great deal to use the standardised form Data2 when carrying out Type 3 SIMPROF tests – either as part of clustering or with **Analyse>SIMPROF** – because permutations are across species within samples, and this will make no sense if species are not first 'relativised' in this way, to total 100% over samples).

So, from the selection in Data1 or from Data2, run **Analyse>SIMPROF>**(Type•Type 2) and take the defaults on the **Next** screen. The output, *MultiPlot2*, contains two graphs, of the real similarity profile (red) and the means and 99% probability limits for that profile under the null hypothesis, and the histogram of absolute deviations π of 999 (further) permuted profiles from that mean, with the real statistic value π indicated by the dotted vertical line. The output is of exactly the same form as previously discussed for single **SIMPROF** runs (see Section 6), and shows with little doubt that there are real species associations to interpret ($p<0.1\%$). With a large number of similarities making up the profile ($50\times49/2 = 1225$), it is inevitable that the probability limits and the real profile will hug the mean curve fairly closely but it is clear that there is an excess of both higher and lower associations than one would expect by chance under the null hypothesis – some of the species are 'positively' associated and some 'negatively' (we retain the terminology of correlations being positive or negative though, as explained in Chapter 7 of CiMC, an index of association defined over (0, 100) is a better measure of species inter-relationships than a correlation coefficient). Note that very few of the 'negative' associations are at the lower limit of IA = 0, which arises when two species are only ever found in different years – this is the result of removing all the low abundance species. [About half the original 111 species were found in three or fewer years – and if you prefer to carry out a species reduction on this type of criterion, you can do so by **Select>Variables>**(In at least n samples where n is), entering that reduced matrix to **Analyse>SIMPROF** and **Wizards> Coherence plots**. Leaving in rare species always results in a tail of fully 'negative' associations.]

v7



Running Type 3 SIMPROF

v7

Type 3 SIMPROF can be run as a single test on the species set defined by the active matrix (which could be under a selection), using **Analyse>SIMPROF>**(Type•Type 3), leading to similar dialog and outputs as for the Type 2 test. However, there are two important differences. Firstly, as noted above, it is now essential that the input matrix is species-standardised, for the permutations to make any sense. Secondly, even if all species in the (reduced) data matrix – Data2 in the above context – are submitted to this single run of Type 3 **SIMPROF**, the procedure is not testing the same null hypothesis as the Type 2 test. That seeks to reject the null hypothesis of no associations amongst any of the pairs of species; Type 3 tests the null that all associations amongst pairs of species in that set are the same. Clearly, it is at least possible (albeit not very likely in practice) for the Type 2 test to reject its null hypothesis, so that there is statistical support for examining the structure of the among-species relationships, but that this further study – which starts with a Type 3 test of all the species – fails to find any species clusters at all. This would happen if all species had exactly the same strong pattern (subject to sampling error) over the samples, e.g. there might be a single strong environmental gradient and all species abundances decline at the same rate along that gradient.

As implied in that comment, a Type 3 test is rarely performed singly. More typically, a series of Type 3 tests are carried out automatically, working down the branches of a hierarchical species clustering, in order to determine the coherent species sets, i.e. all the nodes which give rise to non-significant Type 3 tests. Structure displayed in the species clustering below that point is always in red, indicating that there is no statistical support for interpreting these further sub-divisions. Whilst

v7

Wizards>Coherence plots only performs group average (UPGMA) clustering – though using any selected association/correlation coefficient – the other hierarchical clustering methods offered by PRIMER 7 can be applied to variable resemblances by running **Analyse>Cluster>** and choosing the agglomerative CLUSTER (and an alternative linkage method) or the divisive unconstrained UNCTREE, and specifying (✓SIMPROF test), see Section 6. [Even the constrained LINKTREE routine (Section 13) would run, though defining sensible constraints might prove elusive!].

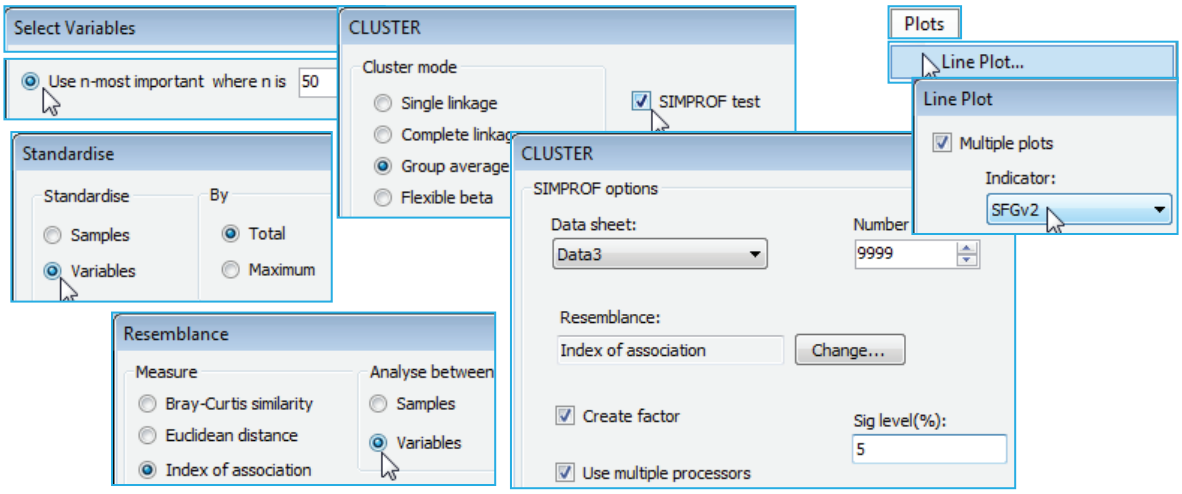
As with the earlier **Matrix display**, it is instructive therefore to recreate the exact steps of **Wizards >Coherence plots** by running the component routines. On data Linnhe macrofauna abundance take **Select>Variables>**(•Use n-most important where n is 50), then run **Pre-treatment>Standardise>** (Standardise•Variables) & (By•Total), **Analyse>Resemblance>**(Measure•Index of association) & (Analyse between•Variables), and **Analyse>Cluster>CLUSTER>**(Cluster mode•Group average)

v7

& (✓SIMPROF test) with the default options of Index of Association, a 5% significance level and (Add indicator named: SFGv2). The resemblance matrix and dendrogram structure will be identical to that for the previous run of **Coherence plots**, but if the SIMPROF groups are not absolutely the same this will be due to the random nature of the permutations and the fact that one of the tests is rather borderline to the 5% significance level, sometimes falling one side and sometimes the other. A larger number of permutations would firm up the true significance level (9999 is preferable if it can run in minutes – actually seconds here) but that does not address the arbitrariness of specifying

v7

a 5% significance level for this (or any!) test. The Somerfield & Clarke (2013) paper suggests re-running at 0.1%, 1%, and 5% to see how stable the final groupings are to the choice of level – the more stringent significance levels will produce a smaller (or the same) number of coherent groups. The final step is to run **Plots>Line Plot** on the (reduced) species-standardised data sheet (probably named Data3), taking (✓Multiple plots)>(Indicator: SFGv2) to recreate the previous *MultiPlot1*.



Line plots *vs* Shade plots

v7

For this (reduced) Loch Linnhe data matrix, e.g. with the selection of 50 species made in Data1, it is straightforward to create also a shade plot. There are, of course, differences in visual impact in the way matrix entries are represented by *y* axes of a line plot or depth of shading but there are also some contrasts in the emphases that PRIMER gives to the two displays:

a) both would usually require a reduced number of species to be analysed and use the same species resemblance measure (index of association) and species clusters. The automatic **Wizards>Matrix display** does not, however, carry out the Type 3 SIMPROF tests of the **Coherence plots** routine – though a direct run of **Shade Plot** could certainly display the groups resulting from these tests;

b) **Wizards>Coherence plots** generates line plots based on the standardised species values, usually without transformation, so that the emphasis is on which species follow the same patterns, or have statistically different responses over samples. In contrast, **Wizards>Matrix display** would usually show the transformed (and not species standardised) values by depth of shading/colouring, since its primary emphasis is on how the species contribute to the multivariate analyses of the samples. It is perfectly possible, however, using a (non-wizard) run of **Shade Plot** to create a display of the same standardised (non-transformed) values as are used in **Coherence plots** – see below;

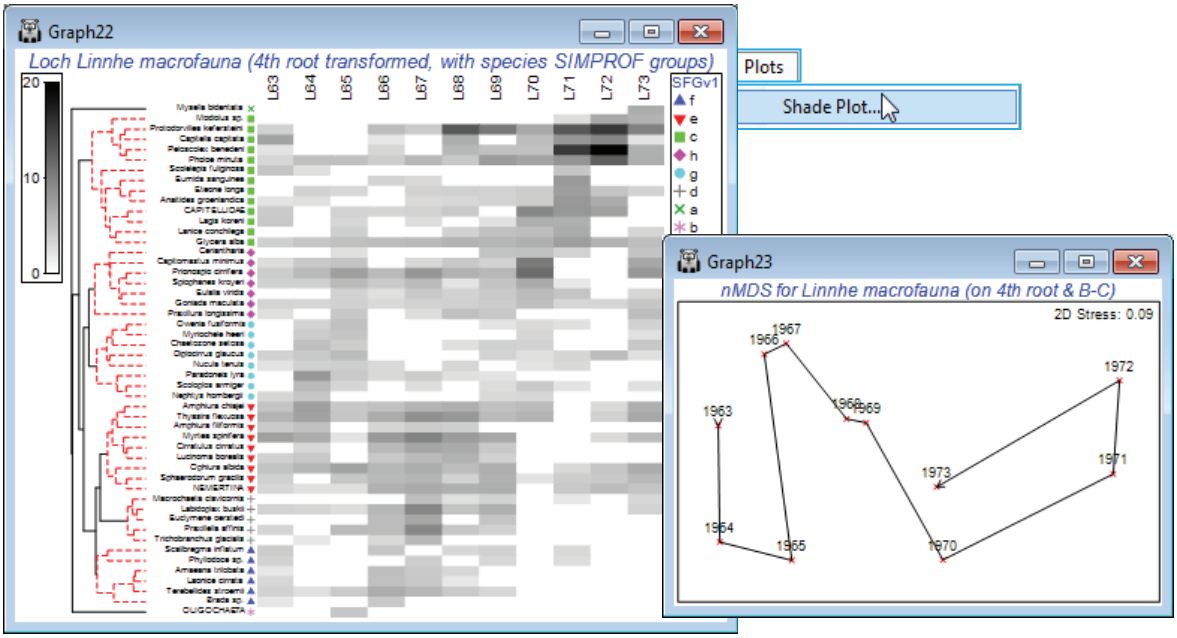
c) **Matrix display** and the direct **Shade Plot** routine will terminate with an error if a worksheet of type *Environmental* is submitted to them – they will only accept data of types *Abundance*, *Biomass* or *Other*. This is because a shade plot is designed to display quantities, with blank (white) space for absence through to black for the (rounded up) largest quantity in the matrix, with the same scale applying to all variables (usually taxa). The measurement units for environmental variables usually differ (or at least the ranges occupied within a common scale differ strongly) and normalisation is required to produce a common scale – zero now has no particular meaning. One could force the software into producing a shade plot for such environmental data – change the data type with **Edit>Properties>**(Data type•Abundance) then add a constant (*k*) to move the scale to positive values, with **Pre-treatment>Transform (individual)>**(Expression: V+*k*) – but the PRIMER shade plots

v7

are visually set up to suit only quantity data. In contrast, the **Coherence plots** and multiple **Line Plot** routines work well for all data types, though again for variables converted to a common scale, such as standardising for taxa and normalising for abiotic data – the latter is also seen below.

v7

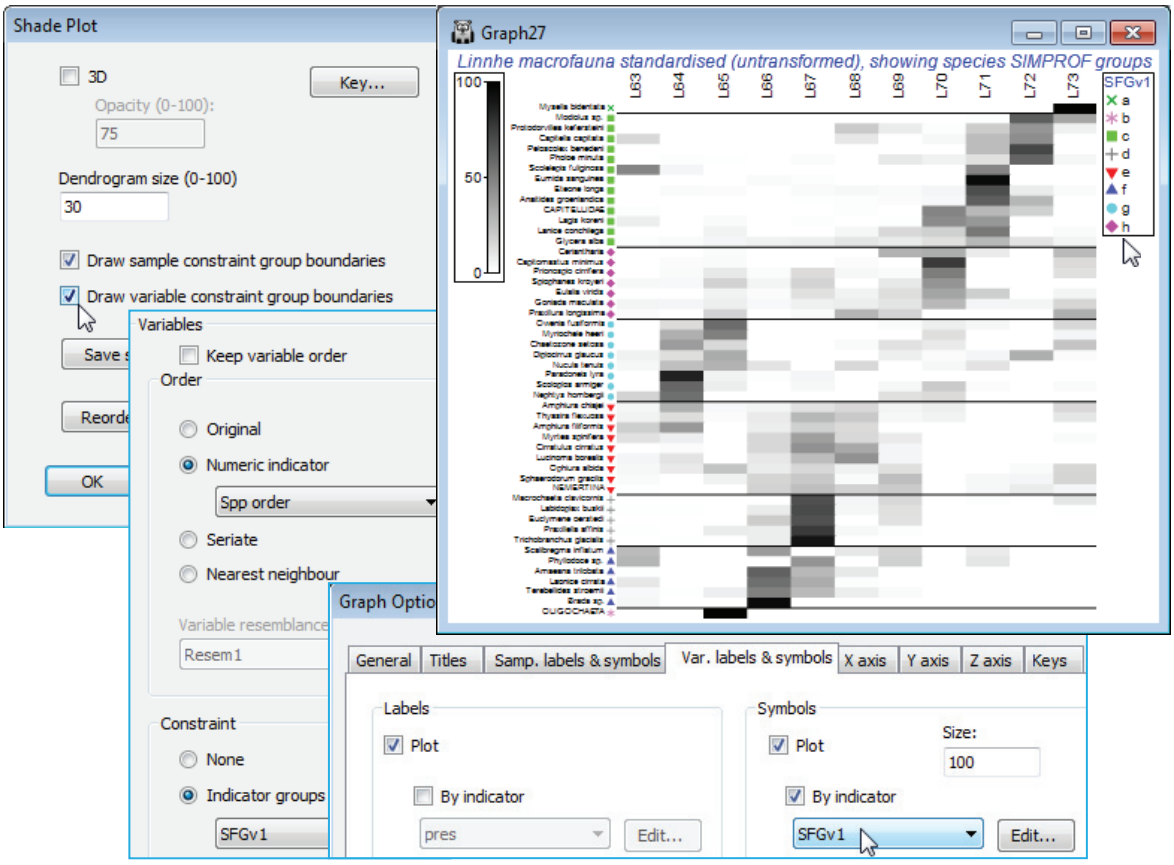
For the selected 50 species from the Linnhe data (Data1), run **Pre-treatment>Transform(overall)>**(Transformation: Fourth root) and **Plots>Shade Plot**. Note that the plot (which is not wonderfully clear!) has the years already in the right order, since that was the sample order in the matrix, but the species need grouping according to the dendrogram given by the run of **Coherence plots**, Graph1, with **Graph>Special>Reorder>**Samples>(Order•Original) & (Constraint•None) and Variables>(Order•Seriate>Variable resemblance Resem1) & (Constraint•Variable dendrogram Graph1) and (No. of seriate restarts: 999). The species SIMPROF groups are seen in red on the dendrogram but can be accentuated by **Graph>Var. Labels & Symbols>**(Symbols✓Plot>✓By indicator SFGv1). This is now a plot of the dominant ~50% of the species contributing to an *n*MDS based on the 4th-root transform of the original Linnhe macrofauna abundance matrix under Bray-Curtis similarity.



Shade plots showing coherent sets & variable boundaries

v7

In order to display a shade plot carrying precisely the same information as the coherent species line plots, first save the species order in the above shade plot – in order to make it easier to compare the two shade plots – using **Graph>Special>Save variable order>**(Indicator name for variable order Spp order), then with the active matrix this time as the species-standardised (but not transformed) sheet Data2, run **Plots>Shade Plot** and note that now many species rows contain the darker entries, i.e. those whose dominant abundances are in a single year, the common scale (0, 100%) now being relative to the total abundance for each species over all years. Rearrange the species in the previous order by **Graph>Special>Reorder>**Variables>(Order•Numeric indicator Spp order)&(Constraint• Indicator groups SFGv1), and after **OK** returns you to the first of the **Special** dialog boxes, take (✓Draw variable constraint group boundaries). This is the first time we have seen horizontal lines on the shade plot, separating the coherent species groups, and it certainly helps the visual message. The plot below also adds symbols for the species groups, using the **Var. Labels & Symbols** tab as previously, since the key does make a link to the earlier line plots of these groups, for which the indicator level (lower case letter) is given in the sub-titles. Remember also that you can re-order the constituents of a key by clicking on the key and using the Move arrows, which has been done here to re-arrange the indicator levels alphabetically. Abbreviations of the species names can sometimes be necessary and this is simply achieved on the **Var. Labels & Symbols** tab again, by putting them into another indicator Spp abbrev and (Labels✓Plot>✓By indicator Spp abbrev).



‘Mondrian’ shade plots, with sample and variable boundaries

v7

A final possibility for this Loch Linnhe data is to display both horizontal and vertical divisions in a shade plot, from sets of Type 3 and Type 1 SIMPROF tests respectively. Though it might normally be preferable to display the years in their chronological order, it could make good sense to examine which years cluster together in terms of their communities, i.e. whether there is some evidence that the more contaminated conditions of the early 1970’s and the ameliorated discharges by the time of the final year are paralleled by changes in the multivariate community structure, as suggested by sample SIMPROF tests. So, on the Bray-Curtis sample resemblances from fourth-root data on all species – used earlier to produce the *n*MDS plot of the 11 years (sheet Resem3 in the Explorer tree shown below) – run a sample clustering with SIMPROF (Graph28), creating factor YrGps, and on any sheet in the branch of the last shade plot Graph27, e.g. Data2, take **Edit>Factors>Import>** (Worksheet: Graph28)>**Select**>(Include: YrGps), so that this factor is now available to that shade plot. On this plot, run **Graph>Special** and both (✓Draw sample constraint group boundaries) and (✓Draw variable constraint group boundaries) should now be ticked, and the **Reorder** dialog needs

v7

v7

Samples>(Order•Seriate>Sample resemblance Resem3) & (Constraint•Factor groups YrGps), with the species selections unchanged at Variables>(Order•Numeric indicator Spp order) & (Constraint• Indicator groups SFGv1). Again increase the seriation restarts for the samples to 999. Returning to the first **Special** dialog box, take (Number of levels: 5) and **Key>**(Colours•Two colours), choosing red and yellow (left and right). Finally, from the shade plot, tidy up the sample labels by **Graph> Sample Labels & Symbols>**(Labels✓Plot>✓By factor Year) and you can remove the key for the species SIMPROF groups species keys, if you wish, by the **Keys** tab on this Graph Options dialog, unchecking (✓Plot key) under (Select key to edit: SFGv1), to give the 'Mondrian'(!) plot below.

Further examples of analyses leading to sets of coherent species curves are given in CiMC, Chapter 7 and Somerfield & Clarke (2013) *J Exp Mar Biol Ecol*, with an interesting contrast seen between a (spatial) case with complete species turnover along the samples axis – the Exe nematode study seen above – and a (temporal) case with more subtle seasonal and impact-related changes in abundance – the Morlaix macrofauna over the Amoco-Cadiz oil-spill period, the MDS plots for which were discussed extensively in Section 8. The latter is especially interesting because it does put together coherent groups of species which: define the seasonal cycles; apparently respond to the oil spill by drastic decrease or increase in numbers, at differing time lags; appear to be completely unaffected; or are perhaps responding to more long-term temporal change. This tool therefore helps put the clear changes seen in a multivariate analysis of samples into their proper biological context of temporal/spatial/experimentally-driven patterns in the main species making up that community.

Coherent sets of abiotic variables (N Sea biomarkers)

Save the workspace as Linnhe ws, close it and re-open the data on a suite of biomarkers ('health' measures) from flounder sampled at five N Sea sites S3, S5, S6, S7, S9, with 10 replicate samples (pools of fish) from each site. The workspace N Sea ws in C:\Examples v7\N Sea biomarkers holds the datasheet N Sea flounder biomarkers, which was last used in Section 9 to establish differences amongst sites in the full suite of the 11 biomarkers. Two of the variables (EROD and LIPID VAC) were square-root transformed with **Pre-treatment>Transformation(individual)** and the full set (rename them Transformed biomarkers) then normalised to a common scale with mean 0, variance 1 (named Normalised biomarkers). It of interest here to test whether there are coherent biomarker sets – those within a set having the same pattern of response across the samples, and among sets having statistically distinguishable outcomes – and to create line plots of those common responses on the normalised scales. Here, the measurements are all matched to the individual replicates, so it is multivariate response patterns across 50 samples (not just a mean response at each of five sites) which is used to correlate the 11 variables. And it is correlation which, for most abiotic variables, is the relevant measure of variable resemblance (see Section 5). This will be Pearson correlation here, though the non-parametric Spearman correlation can be used instead.

v7

Running **Wizards>Coherence plots** on the active matrix Transformed biomarkers, the routine notes that the data type is defined as *Other* and asks for confirmation that you would like to treat these as environmental rather than community-type variables (i.e. for which correlation measures rather than the index of association are relevant). It is important to retain the (✓Scale variables) default option, with (•Normalise) if the (Resemblance: Pearson correlation) option is chosen, even though the Pearson coefficient itself includes a normalising step. This is because the permutation process for the Type 3 SIMPROF test exchanges values across variables, within samples, prior to calculating the resemblance, and this makes no sense if the variables do not have the same scale. Of course, an alternative is to enter **Wizards>Coherence plots** from the Normalised biomarkers sheet, with the (✓Scale variables) step not then required. [If you do this and still take (✓Scale variables), PRIMER will remind you of the fact, but this is only a warning because normalising for a second time changes nothing. Similarly, if you start from Transformed biomarkers and propose to use Spearman correlation then you must take (✓Scale variables•Rank) and (Resemblance: **Change**)> (Measure•Other)>(✓Correlation>Spearman rank correlation). Without the initial ranking, permutation of variables again makes no sense and, much more subtly, do not be lulled into thinking that Pearson correlation will then give you Spearman (since Spearman is just Pearson correlation on ranks), because after permuting ranks over variables, independently for each sample, the ranks no longer add to a constant and the re-ranking implicit in calculating Spearman becomes important].

v7

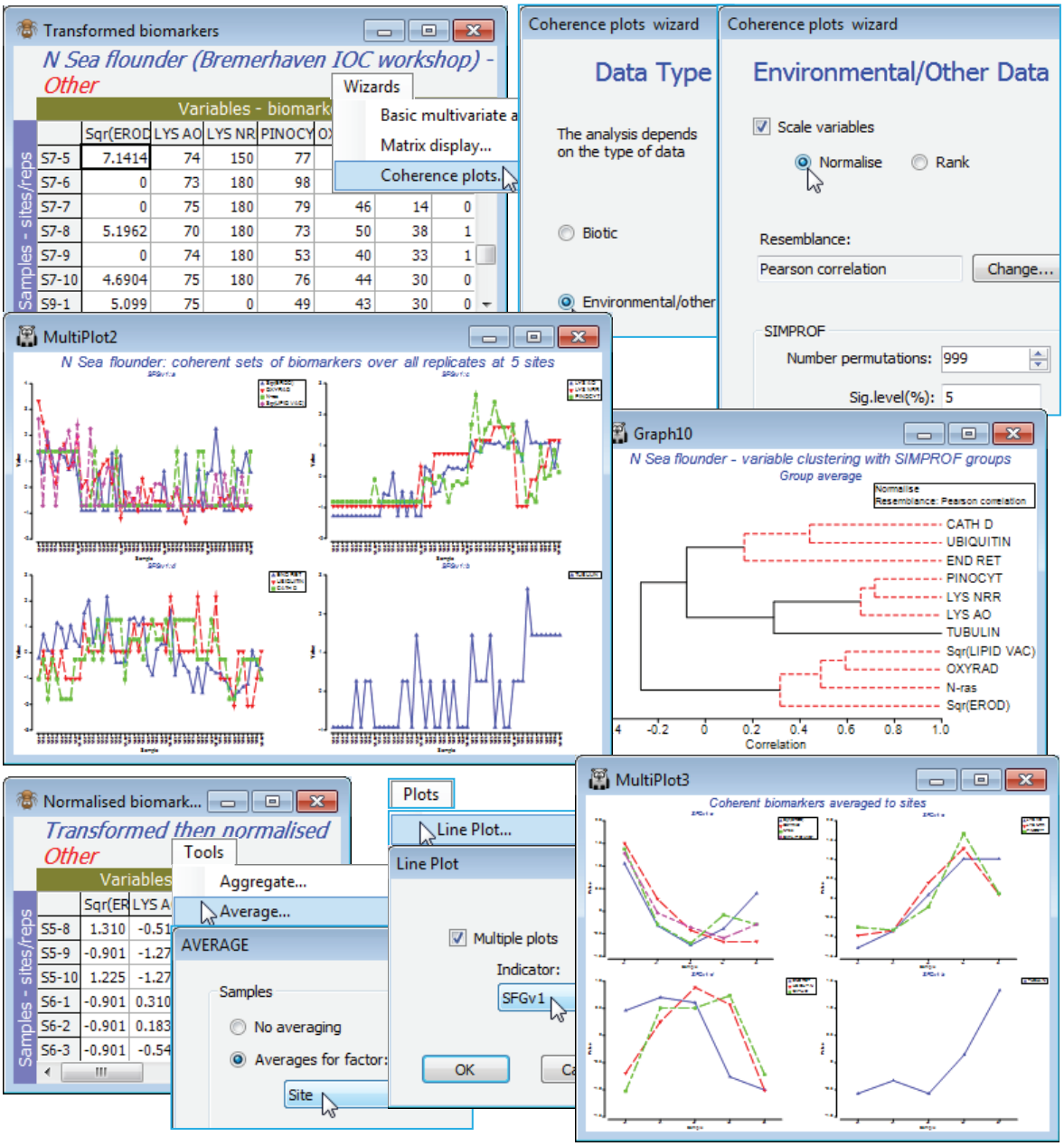
Do not despair at this point! If you take **Wizards>Coherence plots** and the default (•Normalise) and (Resemblance: Pearson correlation) very little can go wrong – that is the point of a Wizard! It is only if you recreate the individual Wizard stages that you must not forget the **Pre-treatment> Normalise variables** step before **Analyse>Resemblance>**(Measure•Other>Pearson correlation) &

v7

(Analyse between•Variables), then **Analyse>Cluster>CLUSTER**>(✓SIMPROF test), taking the default SIMPROF dialog options, and creating (Add indicator named: SFGv1). Look at the cluster dendrogram – whether produced directly or more likely from the Wizard run – and you will see that clustering handles a correlation matrix with negative values. It is not necessary to input only (dis)similarities in (0,100) or distances over (0,∞) to clustering – the dendrogram *y* axis scale can include positive and negative correlations. However, large negative correlations between variables are treated as very low similarity, with only large positive correlations implying high similarity. This may not always be the required behaviour – indeed for these data Somerfield & Clarke (2013) make the case that knowledge of whether a biomarker increases or decreases on impact should be used to reverse some of them, by **Pre-treatment>Transform(individual)>**(Expression: -V), after any other transformation, so that all are expected to decline on impact. (If the behaviour on impact is not monotonic, the index may not be that useful!). That all measurements for those variables are now negative is not important since normalising restores them to the usual range of approximately (-3, 3). The coherent sets may look very different since previous opposite patterns may now match.

v7

The main output automatically from the Wizard – or on the Normalised biomarkers running **Plots> Line Plot>**(✓Multiple plots>Indicator: SFGv1) – is three coherent sets and one singleton. Having used the replicates to determine the variable groupings, a neater summary is given by a means plot: **Tools>Average>**(Averages for factor: Site) on Normalised biomarkers, and redo the **Line Plot**. Save the N Sea ws workspace and close it – it will be needed at the end of the section for SIMPER.



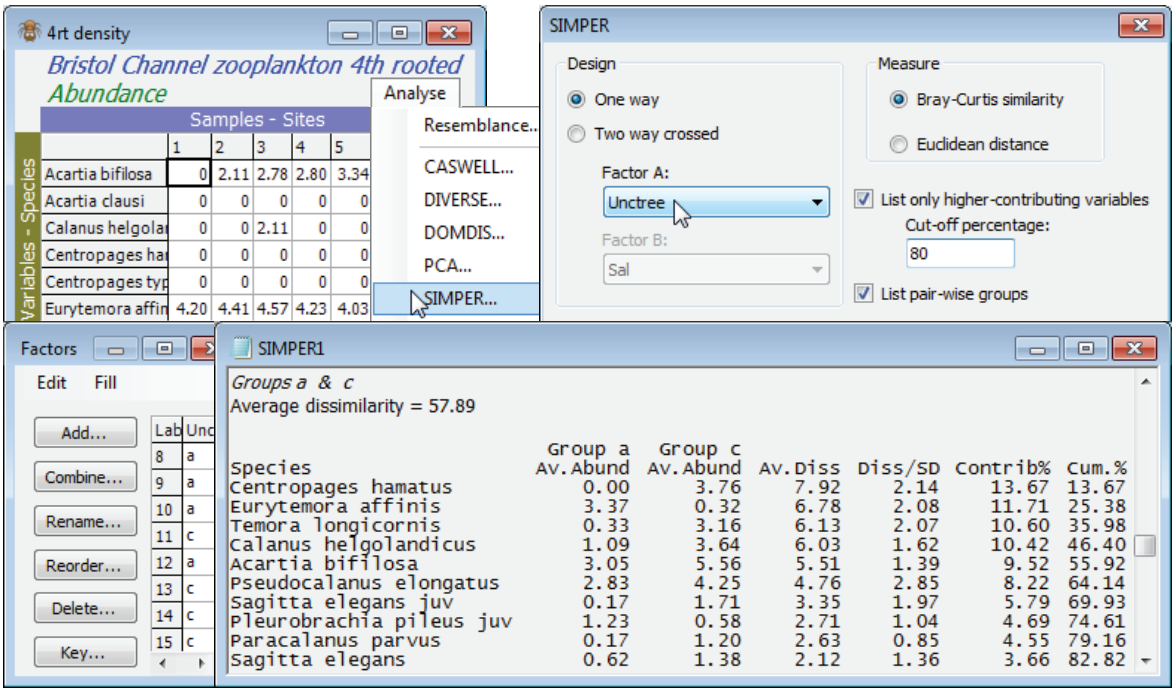
SIMPER (Similarity Percentages)

Shade plots are an excellently succinct way of displaying the abundance (or other quantity) of all the influential species (or other taxon category), in the pre-treated state in which they are input to the multivariate analysis of the community samples. However, they stop short of identifying the precise contribution that each species makes because they take no account of the resemblances calculated on these (transformed) abundances. In the case where there are (a few) established major groupings of the samples, either *a posteriori* by (Type 1) SIMPROF tests or *a priori* by ANOSIM, the average (Bray-Curtis) dissimilarity between each pair of groups can be broken down into the contributions from each species using the similarity percentages routine **Analyse>SIMPER**. This is preceded by a breakdown of the average (Bray-Curtis) similarity within each group into species contributions. Both sets of tables – one table for each group defining the *typical* species and one table for each pair of groups defining the *discriminating* species – have their species ordered into decreasing contributions to the overall average (dis)similarity. SIMPER can be used with shade plots to identify species it may be useful to show on a samples MDS in the form of segmented or simple bubble plots (Section 8). However, as with the SIMPROF and ANOSIM tests, the SIMPER routine operates on the (dis)similarities themselves and not the approximate 2- or 3-d ordination space, so is capable of aiding interpretation for established group structures when these are not adequately represented in low-d (i.e. for high MDS stress). It extends also to cases of (squared) Euclidean distance, e.g. from abiotic analyses, broken down into components from each variable, and to two-way crossed layouts, comparing two groups from factor A, within the strata of factor B.

Species discriminating two groups (Bristol Ch. zooplankton)

v7

Re-open the Bristol Channel ws workspace from C:\Examples v7\BC zooplankton, for which the shade plot was seen a few pages ago. The BC zooplankton density sheet was 4th-root transformed (4rt density) prior to Bray-Curtis calculation (BC resem) and binary divisive clustering (**Analyse> Cluster>UNCTREE**), giving four SIMPROF groups of samples (a to d), factor Unctree. On the transformed 4rt density sheet, run **Analyse>SIMPER>**(Design•One way>Factor A: Unctree) & (Measure•Bray-Curtis similarity) & (✓List only higher-contributing variables>Cut-off percentage: 80) & (✓List pair-wise groups). [This last check box has been added in PRIMER 7 to allow runs on very large numbers of groups where the interest is solely in the species which mainly contribute to the within-group similarities defining each group, rather than cross-group comparisons – there can be sufficiently large numbers of pairs of groups to make SIMPER non-viable unless the cross comparisons are excluded, by unchecking this box]. The restriction to a cut-off (of 80% here) is probably unnecessary when there are only 24 species, but can be useful to avoid long tables, listing all species however small their %contribution to the average dissimilarity between two groups.



Species	Group a Av.Abund	Group c Av.Abund	Av.Diss	Diss/SD	Contrib%	Cum.%
Centropages hamatus	0.00	3.76	7.92	2.14	13.67	13.67
Eurytemora affinis	3.37	0.32	6.78	2.08	11.71	25.38
Temora longicornis	0.33	3.16	6.13	2.07	10.60	35.98
Calanus helgolandicus	1.09	3.64	6.03	1.62	10.42	46.40
Acartia bifilosa	3.05	5.56	5.51	1.39	9.52	55.92
Pseudocalanus elongatus	2.83	4.25	4.76	2.85	8.22	64.14
Sagitta elegans juv	0.17	1.71	3.35	1.97	5.79	69.93
Pleurobrachia pileus juv	1.23	0.58	2.71	1.04	4.69	74.61
Paracalanus parvus	0.17	1.20	2.63	0.85	4.55	79.16
Sagitta elegans	0.62	1.38	2.12	1.36	3.66	82.82

v7

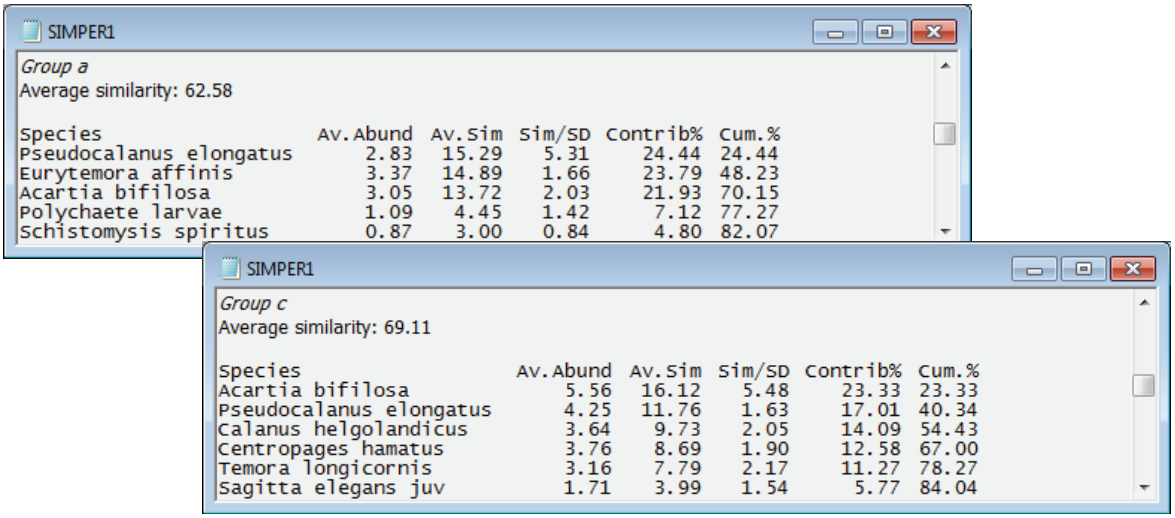
From the results window, *SIMPER1*, find the table comparing *Groups a & c* (above) – these are the groups of sites *a*: *1-10, 12, 24* and *c*: *11,13-22, 25-27, 29* (the left-hand groups in the shade plot in this section). The average of the Bray-Curtis dissimilarities between all pairs of sites (one in *a*, the other in *c*) is 57.9, and this is made up of 7.92 from *C. hamatus,* 6.78 from *E. affinis* etc, given in the third data column of the table. The *C. hamatus* contribution is 13.7% of the total of 57.9 and *E. affinis* gives 11.7% of the total, etc (column 5), and these percentages are cumulated in column 6, until the cut-off of >80% is reached. Column 4 is the ratio of the average contribution (column 3) divided by the standard deviation (SD) of those contributions across all pairs of samples making up this average. A good discriminating species is one which contributes relatively consistently to that distinction for all pairs of sites, i.e. with a low SD and thus a higher ratio (e.g. *P. parvus*, with ratio only 0.85 does contribute something to the difference between *a* and *c* but does so inconsistently). Whether the emphasis is on column 3 – which is used to order the species – or column 4 depends on the context. If you are trying to identify species which contribute the most to the differentiation of those groups in the multivariate analyses then it should be the *Av. Diss.* column – they will tend also to be the species with the larger abundances – but if you are looking for the best indicator of the differences between those conditions the *Diss/SD* ratio should also be considered. It can sometimes pick out species which are completely absent in one group and with very consistent presence in the other, but with low abundance. Columns 1 and 2 aid the interpretation by giving the average abundance (or biomass, cover etc) for each species in each of the two groups – something the shade plot also gives a good feel for (note that both these columns and the shade plot present transformed abundances). *C. hamatus* appears in good numbers in *c*, having been absent in *a*, with the opposite pattern for *E. affinis*. Back-transforming for *C. hamatus* gives a change from 0 to 159 (3.55^4) on the original abundance scale (time-averaged numbers per m^3), with the reverse pattern for *E. affinis*.

v7

v7

Species typifying a group

Earlier in the results window, tables are given of the contributions of each species to the Bray-Curtis similarity *within* each of the groups (see Chapter 7 of the methods manual for the formula).

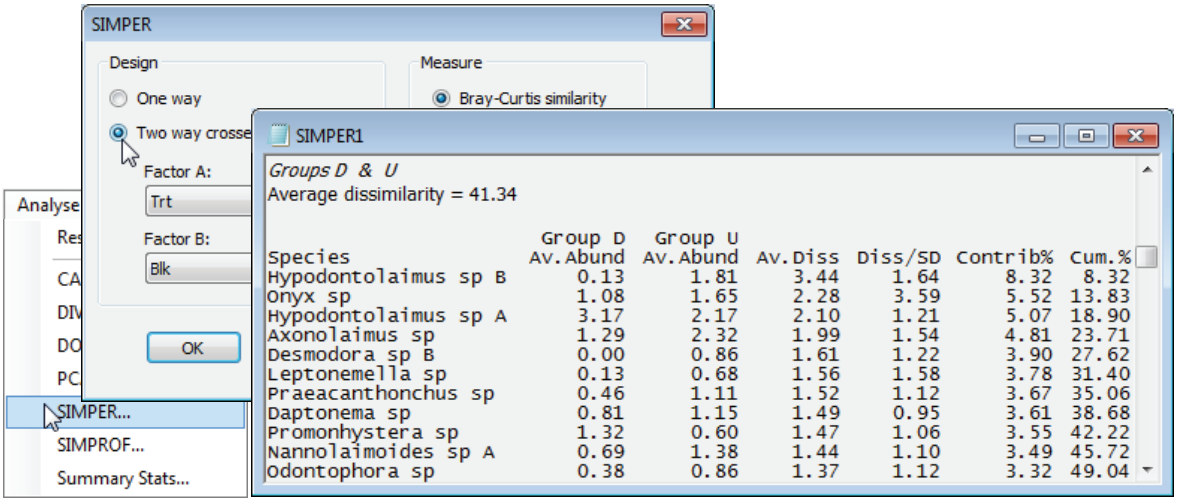


The average Bray-Curtis similarity between all pairs of sites in group *a* is 62.6, made up mainly of contributions from just three species: *P. elongatus* (15.3, i.e. 24.4% of total), *E. affinis* (14.9, i.e. 23.8%), *A. bifilosa* (13.7, i.e. 21.9%), with a cumulative contribution of 70.2% of the total within-group similarity (the list is again truncated at 80%). These species can be described as typical of *Group a* (they also have a consistently large presence because the ratio of their contribution to its SD, across the within-group similarities, is relatively high, notably for *P. elongatus*). However, *A. bifilosa* and *P. elongatus* are also typical of group *c*, which is why they do not head the list of those contributing most to the discrimination between *a* and *c*. They did feature lower down in that list, since both have higher average (transformed) values in *c* than in *a* (column 1and earlier shade plot).

Save and close the Bristol Channel ws workspace, and if you would like to try another example of SIMPER on a 1-way layout – but this time with *a priori* groups – open the WA fish ws workspace examined in Section 9, where ANOSIM tests showed that all pairwise comparisons of (predator) fish species gave significantly different (prey) diets, except two pairs involving *S. robustus*, for which therefore you should not attempt to interpret the SIMPER table of discriminating species.

SIMPER on 2-way crossed layout (Tasmania nematodes)

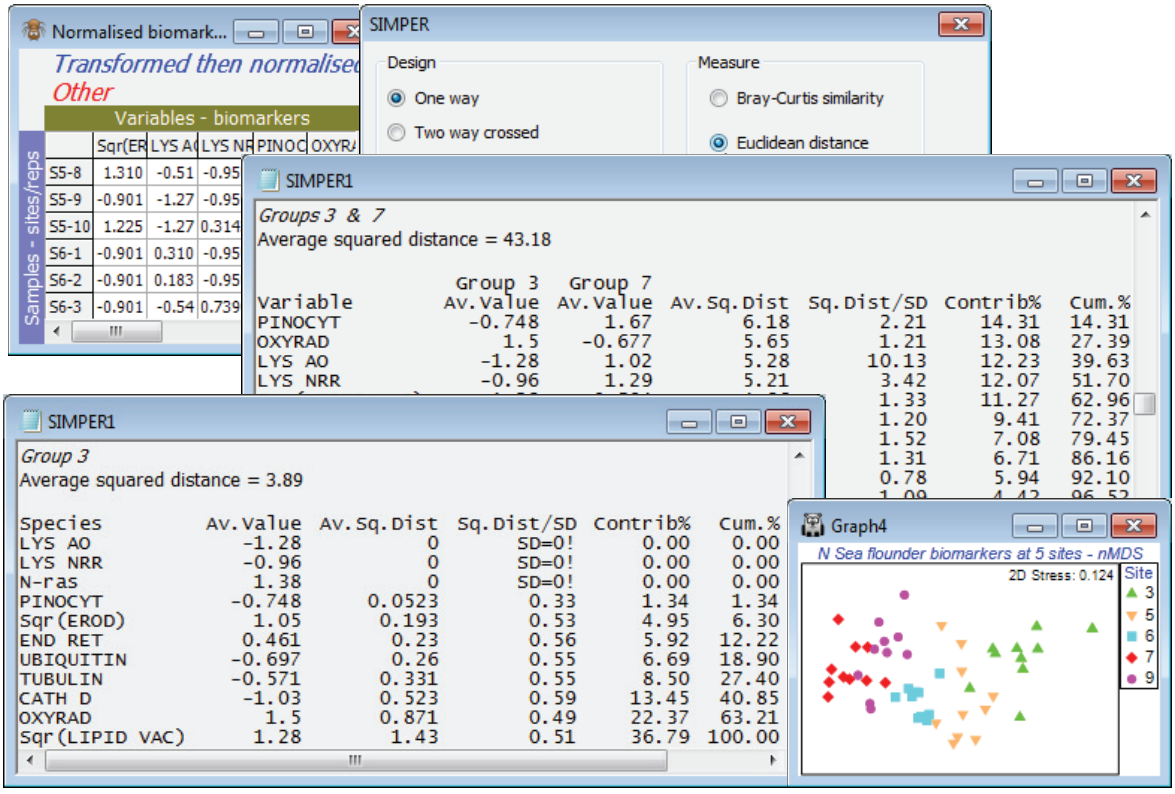
A natural extension to the 1-way SIMPER is to the 2-way crossed design, so re-open the Tasmania ws workspace of meiofaunal communities Tasmania nematodes on a sand-flat, with disturbed and undisturbed patches (factor 'treatment' *Trt*: D or U) at 4 locations (factor 'block' *Blk*: 1 to 4) in C:\Examples v7\Tasmania meiofauna, last seen under 2-way crossed ANOSIM in Section 9. ANOSIM showed there was a clear community difference associated with the natural disturbance by soldier crab activity (not caused by it, necessarily), removing the equally large block differences. In similar fashion, a 2-way SIMPER for treatment differences will remove block differences by looking only at dissimilarities between treatments within blocks and breaking the average of these down into the contributions from species. Starting from the 4th-root transformed nematode sheet gives output:



Of course SIMPER must operate with the active sheet as the data matrix rather than from the Bray-Curtis resemblances since it needs to recalculate all the individual species terms that make up the final dissimilarities. The average dissimilarity between disturbed and undisturbed samples (from the same blocks) is not large, at 41.3, but compares with average similarities within the *D* and *U* groups (only for pairs of samples within each block, thus removing the block effects again) of 68.6 and 73.2 – the header lines of the first two tables of the output – i.e. dissimilarities of 31.4 and 26.8. The *D vs. U* difference is seen to be a sum of small contributions from a rather large set of species. (*Hypodontolaimus sp B* heads the list but this is not because it is the most abundant species overall, e.g. *Hypodontolaimus sp A* has larger densities). In fact, it is unlikely that one or two species will dominate the contribution because of the severity of a fourth root transformation – the heavier the transform, the more species will be involved in the similarity calculation and thus the breakdowns. The symmetry of a 2-way crossed design dictates that the decomposition can be done on the second factor, removing the first, and these tables are also provided, though of less interest in this context.

SIMPER on (squared) Euclidean (N Sea biomarkers)

Save and close Tasmania ws and, as a last example, open the recently closed N Sea ws workspace. On the Normalised biomarkers sheet, run **Analyse>SIMPER>**(Design•One way>Factor A: Site) & (Measure•Euclidean distance), unchecking the box which truncates the listings. The contaminant gradient tends to decline from site *3* (mouth of the Elbe) to *7* and increase on the Dogger Bank (*9*), and the table comparing sites *3* and *7* gives the highest average Euclidean distance squared from all 11 biomarkers – these are also the two endpoint sites of the gradient seen on the biomarker *n*MDS plot of Section 9 (and ANOSIM R = 0.99 for *3 vs 7*). The normalisation puts all biomarkers on an equi-variable scale so all are likely to contribute something, but *Pinocytosis, Oxyradicals* and the two *Lysosomal stability* indices head the list of discriminating variables for these sites (evident also from the *coherent variables* line-plots recently seen). [The SIMPER breakdown is defined naturally in terms of squared Euclidean distance, not Euclidean itself – eqtn (2.13) of CiMC – but this is not important to PRIMER because ANOSIM tests, *n*MDS etc all work on ranks of these distances and those are identical between Euclidean and Euclidean squared]. The starting tables in the output that give breakdown of distances within groups are somewhat less natural than they are for Bray-Curtis, (for which both similarity and dissimilarity can be written as a natural sum over species – see eqtns (7.2) and (7.3), CiMC). They are again read from the top downwards, starting with variables which contribute least to the average Euclidean distance (squared) within a group – the key information to scan being column 2. For site *3*, having a low average within-group distance (squared) of 3.9 (c.f. 43.2 between sites *3* and *7*), the lysosomal stability and *Pinocytosis* are zero, and *N-ras* and *EROD* consistently high for nearly all samples (an indication of impact); these indices fill the top 5 places.



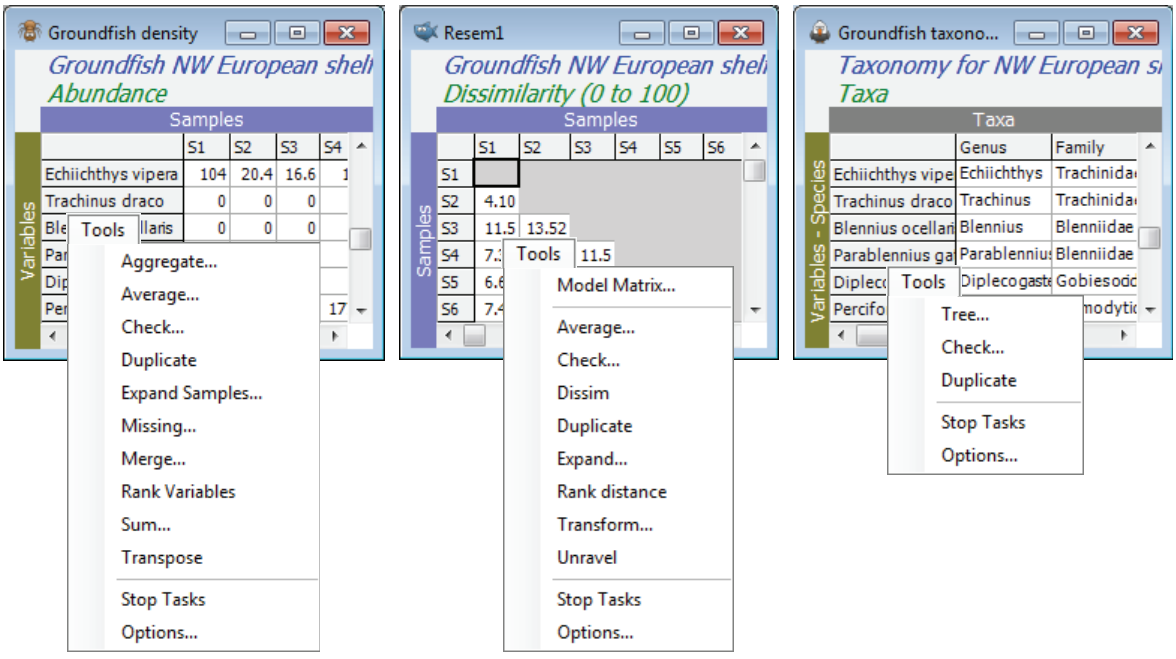
11. General data manipulation (*Tools*, further *Pre-treatment*)

Tools *vs.* Edit menu

Both the **Edit** (see Section 1) and **Tools** main menus carry out 'housekeeping' manipulations on a dataset (or a resemblance or variable information sheet, such as an aggregation file). The operations are usually rather straightforward, and with an obvious outcome, as opposed to the **Analyse** menu which contains the primary statistical routines. The main difference between **Edit** and **Tools** is that items on the main body of the **Tools** menu create a results window, and in most cases also produce a derived sheet of the same type, e.g. a new data sheet from a data sheet. (There are two miscellaneous items at the bottom of the **Tools** menu, **Stop Tasks** and **Options**, which do not fit into these rules, but are there because this is the conventional place for them in Windows applications). Items on the **Edit** menu, on the other hand, never produce a results window and change the entries on the current sheet in some way (sorting labels, inserting/deleting rows or columns, copying and pasting them, defining new factors or indicators associated with the sheet, etc), and do not write the revised matrix to a new window. **Edit** operations on data sheets themselves therefore have a repeated **Undo** option (Section 1), which will back-track through changes you have made to the data sheet entries. **Tools** operations can be re-run, however, perhaps with different options, simply by going back to the previous data sheet – which is always left unchanged, so no Undo facilities are provided. Some **Tools** items apply when the active window is either a data, resemblance or variable information sheet, though with some differences in operation, whereas others are specific to the window type.

v7

Close any open workspace and open Groundfish ws, last seen in Sections 7 and 6, demonstrating cluster analysis. If not available, open the data file Groundfish density in directory C:\Examples v7\Europe\Groundfish, of species counts from 277 samples in 9 sea areas of the NW European shelf (factor area), and also the variable information file Groundfish taxonomy, defining the Linnaean taxonomy of genera, families, orders and classes for the 93 groundfish species monitored. Create a resemblance matrix (Resem1) in any way you like. Now compare the choices on the **Tools** menu when the active window is a data, resemblance or variable information sheet.



The section works through the choices in (very roughly) alphabetic order, with a few transpositions where menu items or data sets are better exemplified in combination. One or two more specialised routines will be deferred until they are needed (e.g. **Tools>Expand** in Section 14) and the **Average** (and **Sum**) options have been met sufficiently often in previous pages only to need an initial recap.

v7

Average and Sum on data matrices

Tools>Average and **Sum** operate in the same way on data sheets. For example, when (Samples• Averages for factor: area) & (Variables•No averaging) is selected, they average (or sum) across all samples with the same level of the specified factor, separately for each variable (species), here creating a derived data sheet of averaged (or totalled) communities for each area, which can be input into the same multivariate analysis options as the original matrix. Averages are taken for the specified factor (not across it), e.g. if the above set of 277 locations (identified by a factor *site*) had

been sampled at several *times*, then **Tools>Average** for factor *site* gives time-averaged site means. All factors in the original matrix are taken across to the new sheet, and factors such as *area* would still be well-defined over the 277 sites. (However, if the averaging had been for *times*, across all the sites, then the *area* column in the Factor sheet would consist only of Undefined! entries, since the averaging has mixed different areas). If the number of sites in each of the 9 areas is balanced then **Average** and **Sum** leads to the same ordination because the sheets differ only by a constant factor – most resemblance measures are unaltered by an overall scale change. If replication is unbalanced, however, then it is unwise to use **Sum**, because the outcome (using Bray-Curtis at least) would be sensitive to the different total abundances from the differing group sizes – **Average** is preferable.

A less common option is, for example, **Tools>Sum>**(Samples•No summing) & (Variables•Sums for indicator: class#) which would retain all 277 samples but total the matrix over the species to give just two new variables, class 1 (Chondrichthyes) and 2 (Osteichthyes). Pooling abundances to higher taxonomic levels is quite a common requirement but this is more naturally achieved with the **Tools>Aggregate** routine, discussed below. It is possible to **Sum** (or **Average**, though that is very unlikely) on both the axes, e.g. (Samples•Sums for factor: area) & (Variables•Sums for indicator: class#) would give a 2×9 sheet of totals of each of the 2 classes in each of the 9 areas.

v7

The main difference between **Tools>Average** or **Sum** and the **Analyse>Summary Stats** routine, new in PRIMER 7, is that the former computes means or sums within groups of samples (and/or variables) whereas **Summary Stats** will calculate these (and several other) summary statistics only over the full set of samples or variables (and in succession, not both at once, if both are required).

Average on resemblance matrices

v7

In PRIMER 7, the averaging facility extends to resemblance matrices: **Tools>Average**>(Factor/indicator for groups: area) takes the average, e.g. for area 1 and 2, of all resemblances between pairs of samples, the first in area 1 and the second in area 2. It does this for all pairs of areas, thus giving a (9 × 9) triangular matrix of area resemblances. (These are the values at the head of each SIMPER table, defining dissimilarities between pairs of groups, which are then broken down into species contributions, Section 10 – but they are now more conveniently held in resemblance form). As the dialog implies, the averaging could also take place on variable resemblances, if groups are defined over those – perhaps coherent species groups from Type 3 SIMPROF tests (Section 10).

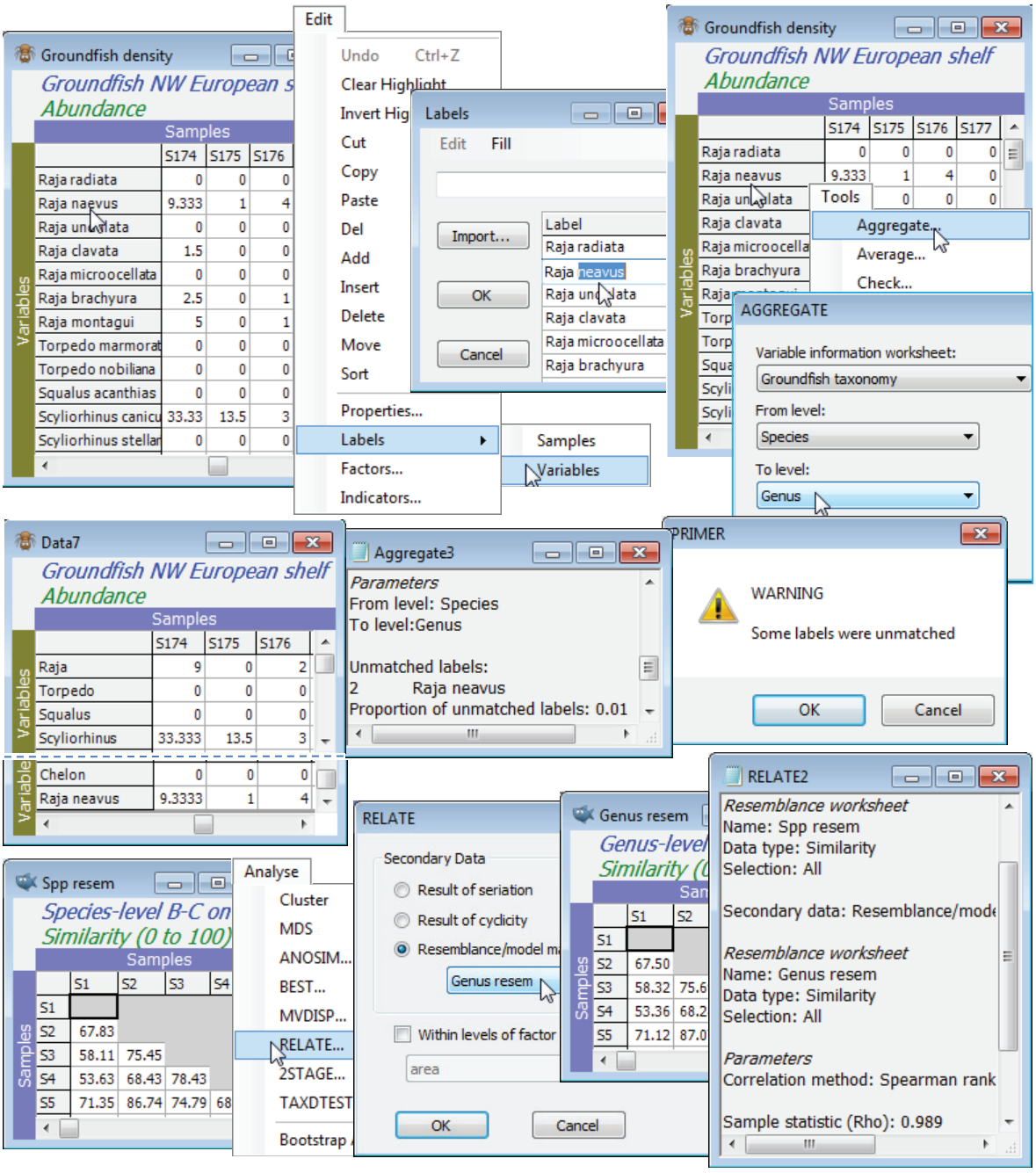
Aggregation

So far we have only seen *variable information* sheets, containing taxonomic (or other) hierarchies (*.agg files), used in calculating specialised forms of resemblance which exploit the relatedness of species in the samples being compared (Section 5). More significantly, this idea of relatedness or distinctness, as expressed in the variable information of the whole taxonomic tree, is the basis of a suite of biodiversity measures (Section 15). But the nomenclature of *aggregation file* (*.agg) comes from the original use of such taxonomies simply to aggregate up an abundance (or other) species matrix to, for example, genus level, i.e. to create a matrix of the abundances that would have been recorded had the species only been identified to a coarser taxonomic accuracy. There are several reasons for wishing to do this, e.g. the taxa might be thought too prone to mis-identification at the species level. Perhaps the data matrix was created over time by several taxonomists with differing expertise in particular taxonomic groups – a 'lowest common denominator' taxonomic level would then certainly lead to a more robust multivariate analysis. Alternatively, the motivation might be resource-driven – if it is possible to establish a putative environmental impact through community change just as clearly with a family-level as a species-level analysis then routine monitoring for that type of impact might be more cost-effectively carried out with data identified to the coarser level. Chapter 10 of CiMC gives many practical comparisons of species- and higher-level analyses.

Whilst, as noted above, pooling the entries for species subsets, separately for each sample, could be accomplished by setting up an indicator and using **Tools>Sum**, this is more conveniently carried out with **Tools>Aggregate**. This works on the original data sheet (prior to any transformation) and specifies a variable information (aggregation) sheet and the hierarchical levels between which the aggregation needs to take place. Of course, unlike data and resemblance matrices the aggregation sheet is not restricted to numeric entries – its variable labels will typically be full species binomial names, and the subsequent columns the increasingly higher level (genus, family, order etc) names. The advantage of pooling using **Aggregate** is that the variable information file of the taxonomic (or other) hierarchy can be a *look-up table* which applies to a wide range of different data sets. There is no necessity for it to have the same number of species, or for those species to be in the same order,

as in the data matrix, as long as all the data matrix species can be found in the more comprehensive faunal list which constitutes the aggregation sheet. Correct (or at least consistent!) spelling is thus essential, including spaces, periods etc. If a species name is not found, a warning is displayed, the results window lists which names were not matched, and these species are retained – with the same values – and with their species name being the higher-level variable name in the aggregated matrix.

Groundfish density and Groundfish taxonomy should be open in the current workspace. In this case the two sheets have the same full list of 93 species in the same order. With Groundfish density as the active window, **Tools>Aggregate>**(Variable information worksheet: Groundfish taxonomy) & (From level: Species) & (To level: Genus), pools the densities to a sheet which you should rename Groundfish genera. Square-root transform both data sheets and compute Bray-Curtis similarities. There is little point in trying to compare the *n*MDS ordinations for the two cases since the large number of samples (277) makes 2- (or 3-d) representations inadequate (high stress). But Sections 13 & 14 make much use of the idea of non-parametric correlation of resemblance matrices, e.g. with the **Analyse>RELATE** routine giving a measure of agreement in representation of sample relationships. Running this on the species and genus similarities gives a high level of agreement, ρ=0.989. You might like to start the example by mis-spelling a species name (e.g. Raja neavus) to observe the consequences, then change it back before running the comparison.



Check on aggregation files

Use the open aggregation file, Groundfish taxonomy, to show the smaller set of **Tools** items (**Tree**, **Check**, **Duplicate**) available when the active window is of *variable information*. **Tools>Duplicate** has been seen previously for worksheets and plots (in Sections 3 and 8). Here it has the same effect, taking a copy of the Groundfish taxonomy window, called Vinf1, to the head of a fresh branch in the Explorer tree. Insert the following errors in Vinf1 to demonstrate the **Tools>Check** option:

a) overwrite Raja clavata (row 4) in the Species column with Raja radiata (by taking **Edit>Labels> Variables** and double clicking in the Raja clavata label and typing in the incorrect name);

b) whilst in the Labels dialog change Squalus acanthias (row 10) to Squaliformes (note that upper or lower case does not make a difference when matching names), and **OK** to exit back to the Vinf1 sheet, then delete Squalus and Squalidae from the genus and family name for that taxon;

c) change Rajidae to Torpedinidae as the family name for Raja naevus (row 2).

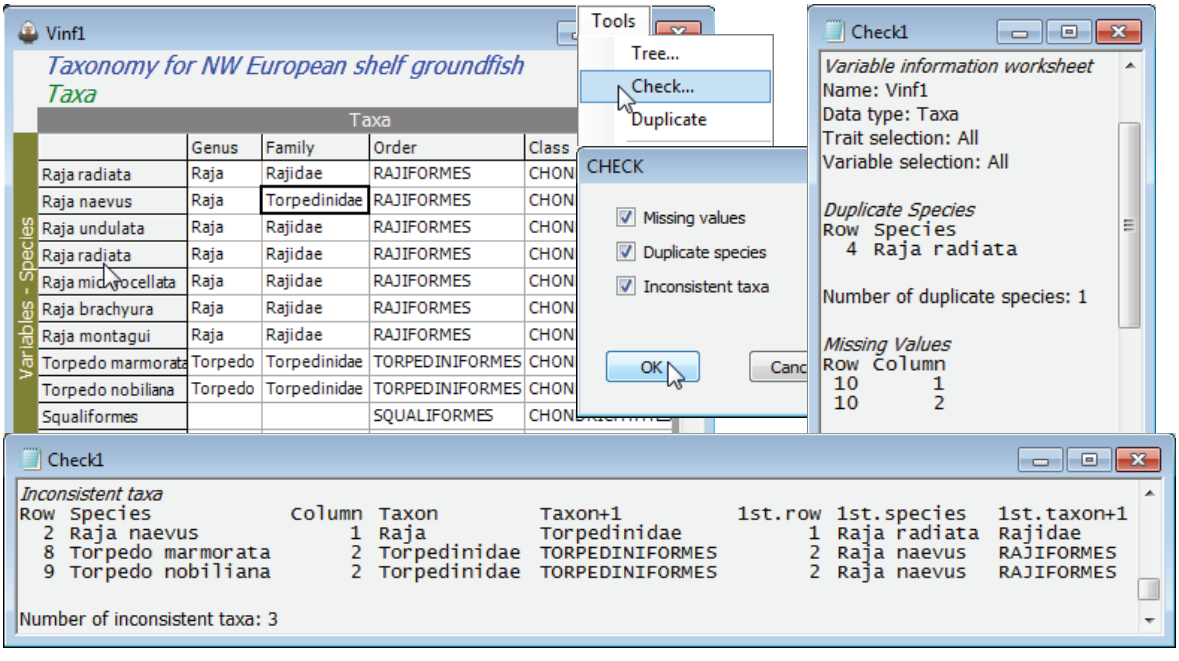
(Note that the row/column numbers of an entry can be found by clicking on it – the status bar at the bottom right displays the current cursor position). Then **Tools>Check** finds three types or error:

1) *Duplicate Species* in row 4 (the repeat of Raja radiata) – labels (samples or variables) should always be unique in a PRIMER worksheet, otherwise matching conflicts can easily result;

2) *Missing Values* (blanks) in row 10. This represents a common situation where only coarser-scale identifications can be made for some taxa. Nonetheless, aggregation sheets need to be complete, in order to avoid incorrect matching. E.g. another species from a completely different order but with a blank family (and genus) entry would be pooled with the *Squaliformes* abundance when the matrix is aggregated to genus or family level, because both entries have the same (blank) family name. Similar problems would occur with taxonomic distinctness calculations (Section 15). So blank entries should be filled with the names from the immediate right or left, depending on the context (often it make sense to fill from right to left). Here put Squaliformes in the two blanks – the routine does not object to the same name being used in different taxonomic levels.

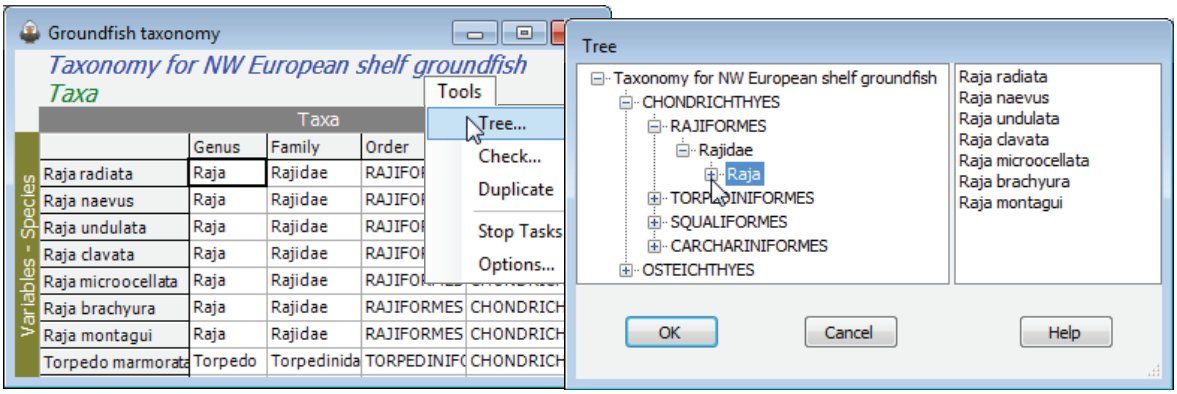
3) *Inconsistent taxa* in rows 2, 8 and 9. In fact there is only one mistake, the family identification of *Raja naevus*, picked up in the correct row (2) because *Raja* has been established by row 1 to be a genus name in the Rajidae family, thus cannot also be a genus name in the family *Torpedinidae*. Quite often, however, an error is not discovered until a conflict occurs much later in the sheet, on a row which may be correct. This is seen in the *inconsistent* identification of the two *Torpedo* genera though neither are wrong. PRIMER 7 has greatly improved its diagnostics here, by listing not just the row and column on which the conflict occurred (in the first 5 columns of the output: row, species name, column, entry in that column, entry in the following column) but also what the conflict with an earlier row was (in the final three columns of the output: the earlier row, its species name, the entry in the earlier row causing the current conflict). So, *Torpedo marmorata* (row 8) in family *Torpedinidae* cannot be in order *Torpediniformes* because in a previous row (it identifies row 2) the family *Torpedinidae* were given as in order *Rajiformes*. With this level of diagnostics, errors in aggregation files (they commonly occur!) should be more easily fixed.

v7



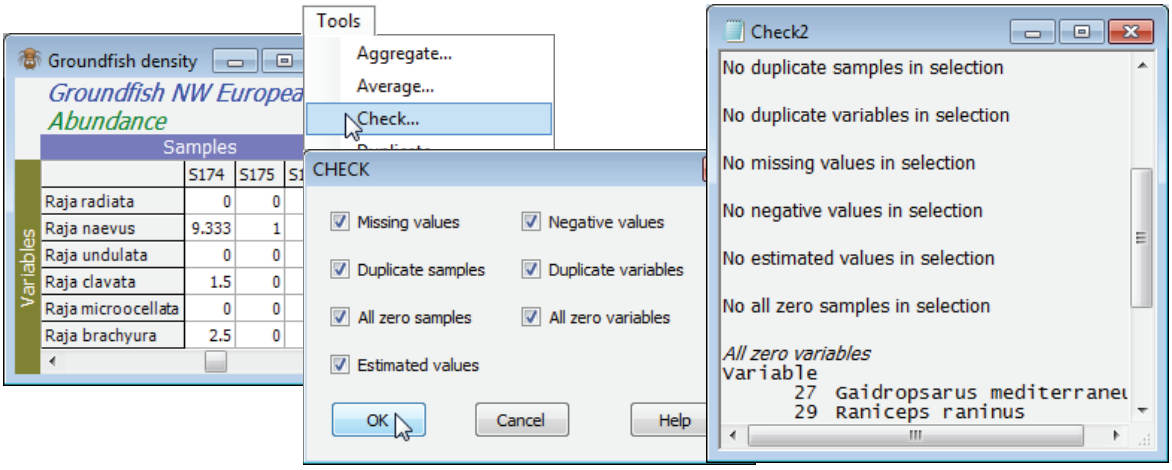
Tree menu

The other **Tools** menu item for aggregation sheets is distinctive to this case, namely **Tools>Tree**; it simply displays the hierarchical structure of an aggregation file in the same way as the Explorer tree, in a left-hand panel. Successive clicking on the ⊞ icons unroll the taxonomic structure, and it can be rolled back with ⊟. No operations can be performed on the display in this state.



Check on datasheets & resemblances

When the active window is a datasheet, **Tools>Check** can check for the following: a) ✓Missing values, identified in the sheet by 'Missing!', and which might have been read in as blank cells in an Excel worksheet for example; b) ✓Negative values, which are not appropriate for abundance-type data analysed by Bray-Curtis, though common for environmental variables (especially normalised) input to Euclidean distance; c) ✓Duplicate sample (and/or) variable labels, which are tolerated for some analyses (warnings are usually given) but are best avoided wherever possible; d) ✓All zero samples (and/or) variables; and e) ✓Estimated values, displayed in red type in the matrix. The latter come from applying **Tools>Missing** (seen shortly) to environmental variables – or to other normally distributed data – containing Missing! cells, which otherwise might not be tolerated by some analysis routines requiring complete data. All or any of the 7 boxes can be ticked. Whether it is important to check for a particular attribute depends on the analysis. For example, species which are zero over all samples will be ignored when Bray-Curtis similarity is computed among samples, and can safely be left in the matrix, but all-zero samples are potentially more of a problem since Bray-Curtis similarity between two blank samples is set to 'Undefined!'. Dependent on the context, these samples might best be omitted, or a different similarity used (e.g. zero-adjusted Bray-Curtis, Section 5), or the entry left as 'Undefined!', i.e. treated as unknown.



Undefined resemblances

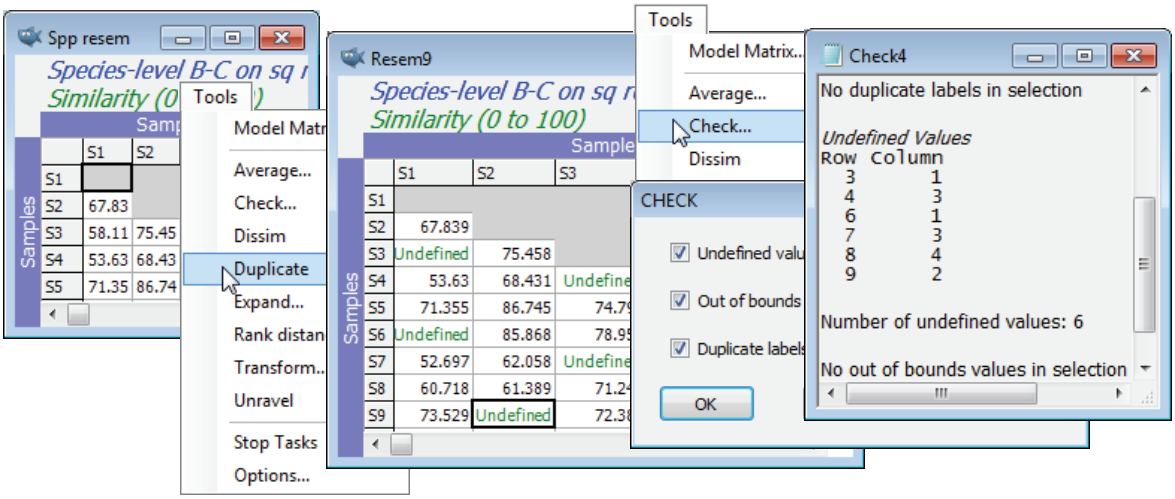
When the active window is a resemblance sheet, **Tools>Check** looks for only three data attributes: a) ✓Undefined values, arising as suggested above; b) ✓Out of bounds values, for distance coefficients (or transformations) that return very large or small values (NaN); and c) ✓Duplicate labels, as above. Blanking a cell in a resemblance matrix sets it to Undefined! status, and several of the core routines using resemblances (e.g. MDS, Cluster, ANOSIM) are carefully written in PRIMER to tolerate a few Undefined! entries, treating them as unknown. (You can appreciate that knowing the similarities S_{12}, S_{13}, S_{14}, S_{23}, S_{24} might enable you to place four samples in relation to each other without knowing similarity S_{34}). Blanking out NaN (Not a Number) entries, to Undefined!, is one possibility therefore, but others may be equally good or better (replacing by a large, but finite value, modifying the coefficient or transformation which generated them etc).

Duplicate

Tools>Duplicate operates in the same way whether the active window is a data array, resemblance matrix, variable information sheet or plot. In the case of a Graph window, **Duplicate** is the only specific option offered on the **Tools** menu, and there are no choices at all for results windows (since they are not capable of amendment once written) except for the **Stop Tasks** and **Options** items which are available on the **Tools** menu whatever the active window. Unless the window is at the top level of a branch already – as a variable information (aggregation) file will always be – an option is offered of (•On existing branch) or (•Start new branch), so that the original links to other sheets and factors can either be retained or a fresh start made. On a new branch, any subsequent amendments to factors, for example, will not then carry back to the originally linked sheets (unless specifically imported by them, with **Edit>Factors>Import**).

v7

On the above resemblance matrix Spp resem for the Groundfish data, take **Tools>Duplicate**> (•Start new branch) and, in the copy, blank out entries at random. A run of **Tools>Check** picks up those now Undefined! entries, but MDS will accept the matrix in this form and produce a plot probably very similar to an MDS run on the intact matrix. Save and close Groundfish ws.



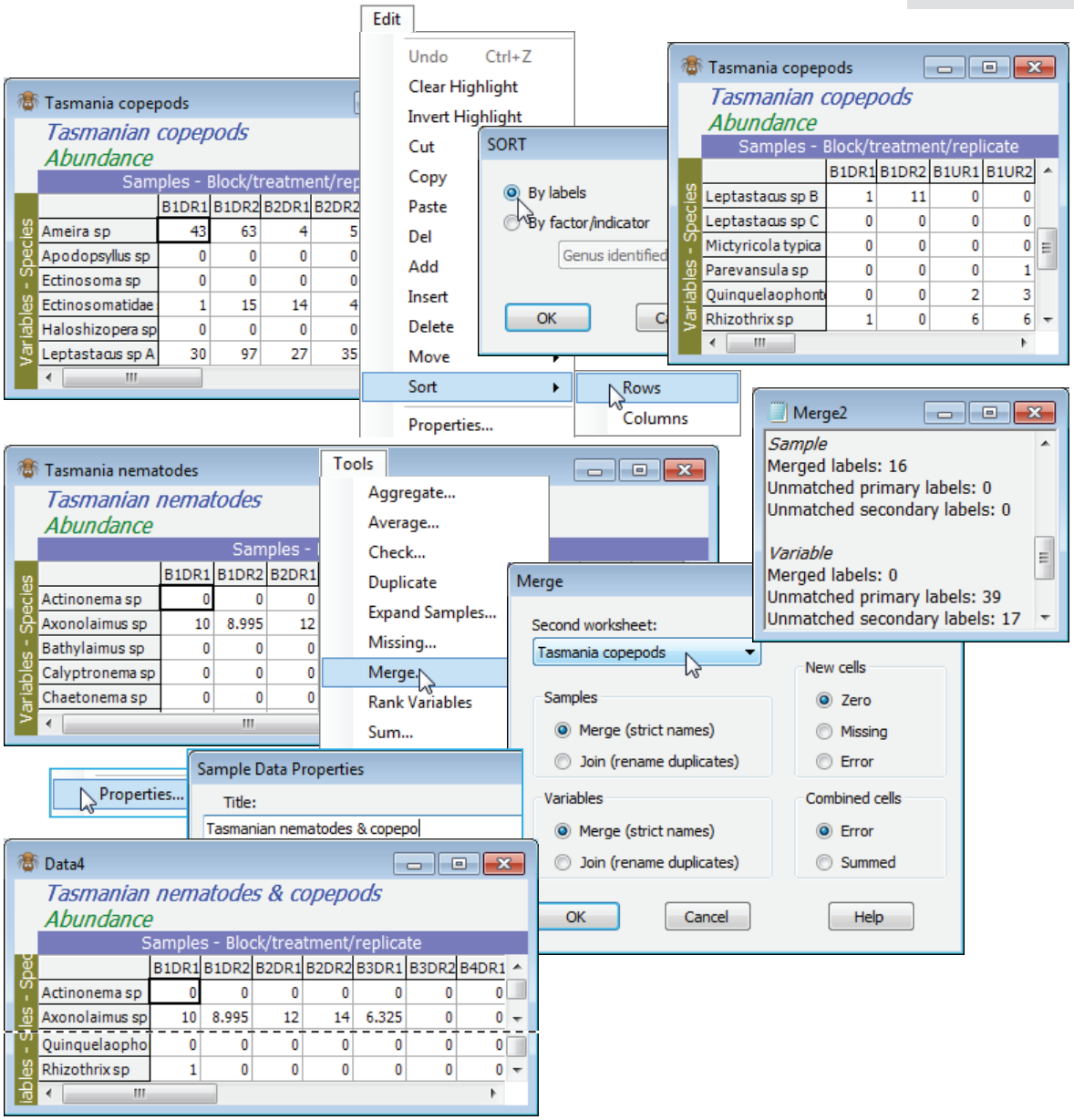
Merge (/join) operations

The **Tools>Merge** menu allows a range of merge operations on two rectangular data sheets. For example, two matrices whose rows are of different variable sets (faunal and algal species perhaps) but with the same sample labels, are automatically joined end-to-end by **Tools>Merge**, with the upper half as the active sheet and the lower half supplied in the (Second worksheet:) box in the Merge dialog. Similarly, two sheets with the same variable labels (species as rows again) but with different sample labels – perhaps the same set of study sites in different years – will be placed side-by-side. The label sets which are in common (at least in part) between the two arrays, and therefore merged in this way, need not appear in the same order in the two arrays – it is the precise label matching which determines the outcome so, as always, consistent spelling is essential.

(Tasmanian meiofauna)

The nematode and copepod datasheets from 16 samples at a Tasmanian sand-flat (C:\Examples v7\ Tasmania meiofauna) were seen in both the previous two sections, in workspace Tasmania ws, but if the latter is not available open Tasmania nematodes and Tasmania copepods in a new workspace. With Tasmania nematodes as the active window, run **Tools>Merge>**(Second worksheet: Tasmania copepods)&(Samples•Merge(strict names))&(Variables•Merge (strict names))&(New cells•Zero) &(Combined cells•Error), i.e. all the default options. The latter two options of new or combined cells do not come into play here, but are discussed later. The resulting merged datasheet now has 56 rows (the 39 nematode species then the 17 copepods). The results window shows that all the samples matched (and the species did not) in the way expected. The title for the new sheet is taken from the first (active) window, so to avoid confusion should be changed using **Edit>Properties**.

Now, re-order the columns in Tasmania copepods by **Edit>Sort>Columns>**(•By labels), sorting the samples in a different (alphabetic) order for the copepod matrix of B1DR1, B1DR2, B1UR1, B1UR2, … than the nematode sample order of B1DR1, B1DR2, B2DR1, B2DR2, …. Nonetheless, a re-run of **Tools>Merge** with Tasmania nematodes as the active window, and with exactly the same options, will result in a merged datasheet identical to the previous one, the ordering of samples having been taken from the first (active) window.

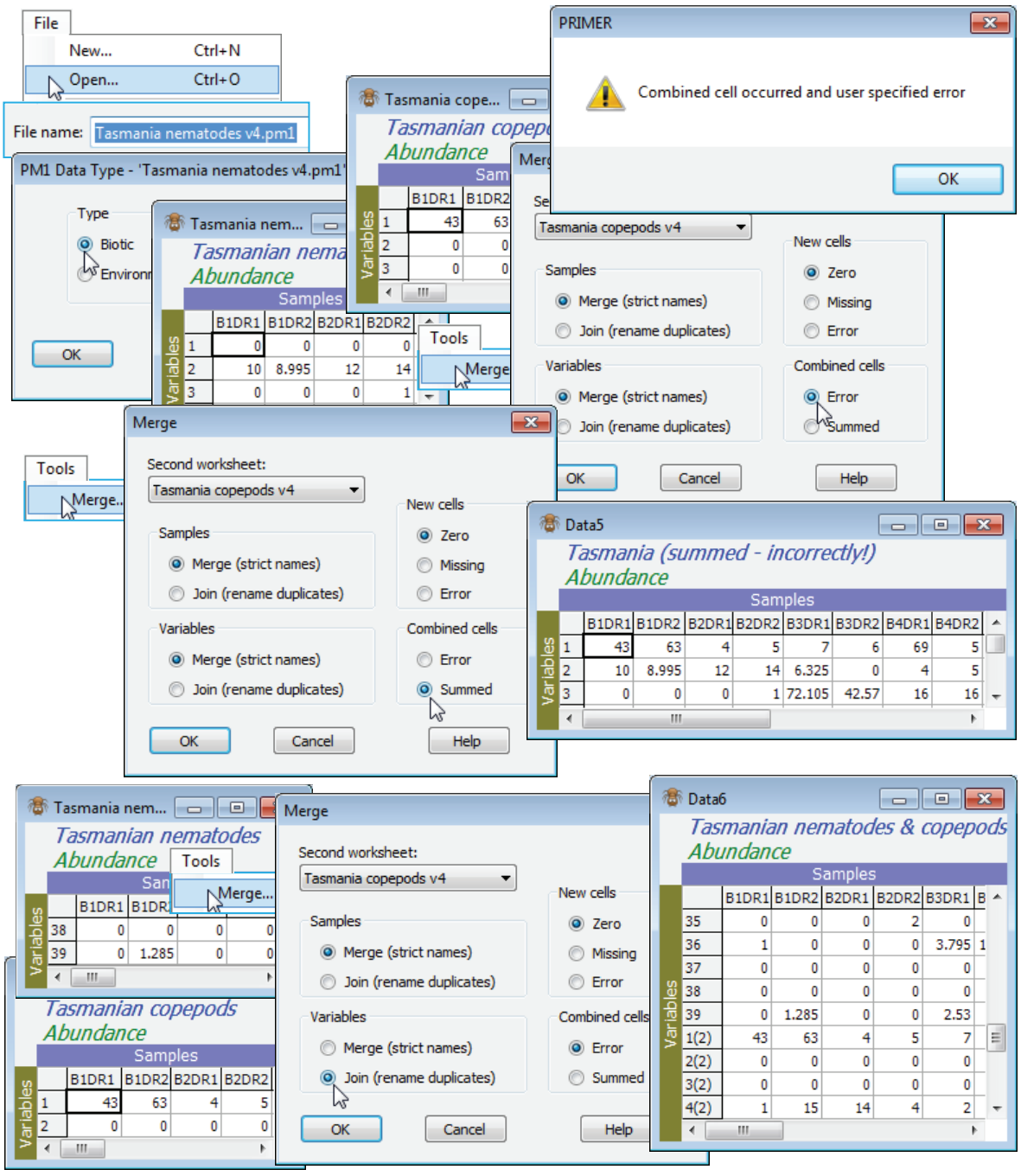


Combined cells in Merge

Occasionally, use of strict label names does not give the this desired outcome, and the default behaviour can be changed to force PRIMER to consider an identical label, but in a different matrix, to be treated as a different name. For example, this might be needed when species names have not been provided for either set, and the variable labels are just the numbers 1, 2, 3, …. Species 1 in the first set is not to be taken as the same variable as species 1 in the second set, and the default options in **Merge** will cause difficulty in this case. Equally possible is the opposite case where the species names match in the two matrices, but the same sample labels are repeated, though should not be equated. Samples collected in year 1 might be labelled by their site identification. A second matrix of data from those sites in year 2 might use exactly the same set of sample labels, i.e. without reference to the year. This causes no confusion if the matrices are to be analysed separately, but a **Merge** under the default of strict name-matching would place the two matrices on top of each other (because they have exactly the same row and column labels!). The two options given in such a case are (Combined cells•Summed) or (Combined cells•Error). The first literally adds the two matrices, element by element. Very often though, this is not the desired behaviour, so the default is the second option: if a **Merge** instruction results in an attempt to combine two cells, an error results.

v7 In the same workspace, take **File>Open** on the data files Tasmania nematodes v4 and Tasmania copepods v4 (in *.pm1 format, from the old DOS-based PRIMER4), which should be read in as Type•Biotic. These are the same nematode and copepod matrices as their *.pri counterparts except that PRIMER v4 held species lists as separate files so both the *.pm1 files have variables numbered just 1, 2, 3, …, though the species are different in the two matrices. A **Tools>Merge** on them, with the default of (Variables•Merge (strict names)) will potentially give combined cells. Try this with

both (Combined cells•Error) in place, to note the error message and the fact that execution then stops. Then repeat with (Combined cells•Summed) – those cells with the same species and sample numbers *are* then simply added together. This may occasionally be a useful option, e.g. it would allow for easy collation of data for the same samples by several different observers (though it must be debatable whether such a piecemeal approach to data matrix construction – losing information on potential observer differences – is often desirable). Taking nematode species 1 to be the same taxon as copepod species 1 and adding the two counts is clearly nonsense in this context, however. The solution, if it is easier to join the arrays in PRIMER and then rename the variable labels later, is next described (the Join option) – this forces the arrays to be placed one after the other.



Avoiding strict label matching

The best policy to avoid confusion is to use precise, unique species and sample labels (typically, the sample label would be a conglomeration of all the different study design factors and a replicate number). However, conflicting desirable criteria can sometimes arise, e.g. when the pattern of sites from year 1 is to be compared with the pattern in year 2, using the RELATE test (Section 14) on the two separate similarity matrices, *identical* sample (site) labels are ideally needed in both arrays, so they can be matched. But, as just pointed out, a Merge of the two data sheets underlying these similarities (so that both year 1 and 2 sites can be seen on the same *n*MDS say) requires the sample

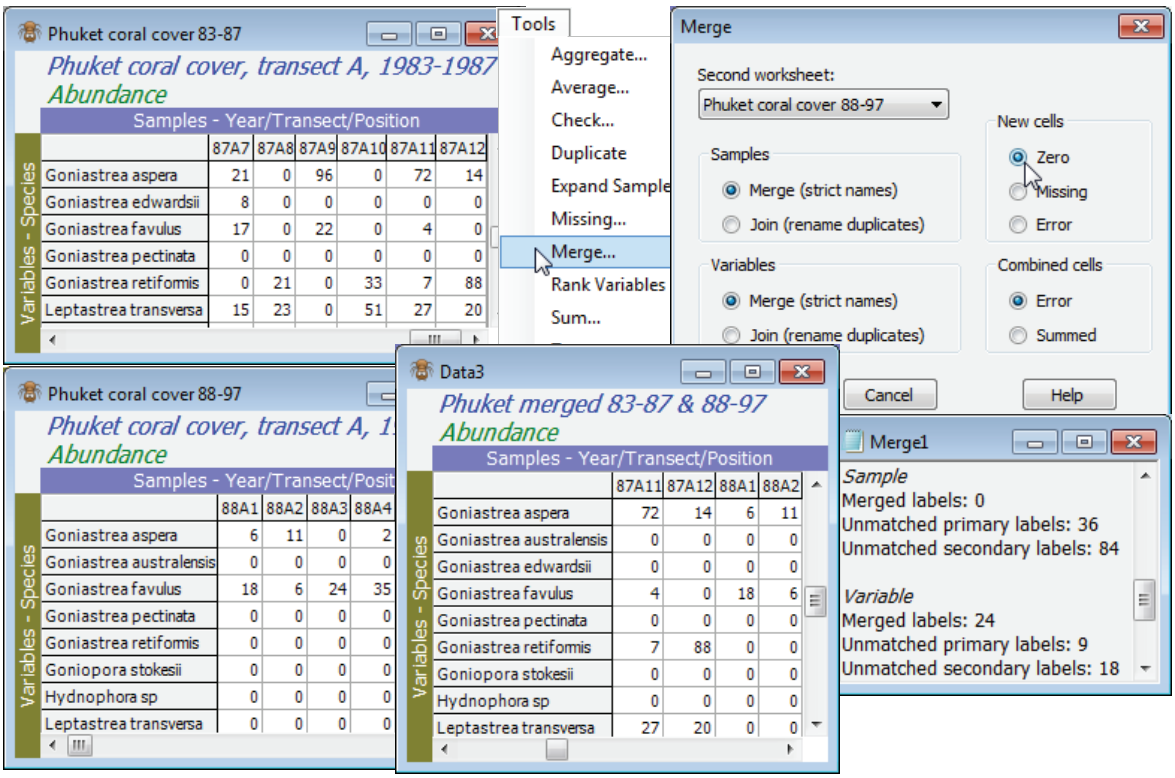
labels to be *different*. Thus, PRIMER is not dogmatic about label matching: several routines, which include **Merge** and **RELATE,** are able to 'fudge' the matching and provide a natural alternative. In **Merge**, this is shown above, using the •Join (rename duplicates) option, used either for Samples or Variables (or possibly both, to create a block diagonal matrix, though this is unlikely to be needed). For Tasmania nematodes v4 and Tasmania copepods v4 sheets to be placed one under the other, even though they share species labels, take **Tools>Merge>**(Variables•Join(rename duplicates)) and defaults for the other options, i.e. (Samples•Merge(strict names)), and there should be no combined or new cells. The copepods are labelled 1(2), 2(2), …, to distinguish them from nematodes 1, 2, … Save the workspace Tasmania ws and close it.

Merging non-uniform species lists

Perhaps the greatest benefit of the strict label matching in PRIMER is the ability to **Tools>Merge** assemblage data when two sets of samples, taken at different times or places, are not recorded on a common data sheet, with predetermined taxonomic categories. Species names, or other operational taxonomic units, must be consistently spelt (even to spaces) in the separate lists, so that the strict matching of variable names can take place. But there is then no necessity that the two sheets hold the same set of species, in the same order. Typically, lists will be of different length, with some species in each list not appearing in the other. Using (•Merge (strict names)) copes automatically with this, filling any spaces created in the merged array either with (New cells•Zero), relevant for assemblage-type data, or with (New cells•Missing), more appropriate for environmental variables. A third option (New cells•Error) stops the procedure with an error message if any new cells are created. This can be a useful safeguard if the intention was to join two data sheets with exactly the same set of variables – an error alerts you to the fact that there may be variable names misspelt.

(Phuket coral reefs)

The Ko Phuket coral reef assemblage data was introduced in Section 8 and the workspace Phuket ws last seen in Section 9. In each sampling year, 12 plotless line-samples were taken along a fixed onshore-offshore transect (A) and area cover determined of each coral taxon. From the directory C:\Examples v7\Phuket corals you will need to have open the three *.pri files of data for different runs of sampling years: Phuket coral cover 83-87, 88-97 and 98-00,only the first two of which were opened in earlier sections. (The early years straddle sedimentation impact from dredging operations for a new deep-water port, 1986/7, and the later ones a sustained Indian Ocean high pressure period with desiccation from lowered sea levels, 1998, with a more stable environment in between). Note the different (but overlapping) species lists of these three sheets. With the active matrix of 83-87, **Merge** this with 88-97, and merge the result again with 98-00, choosing zeros for the new cells, and tidying up the new sheet appropriately (e.g. renaming the window, amending the title with **Edit>Properties** and sorting the species in the merged sheet with **Edit>Sort>Rows>•**By labels).



(Clyde dump-ground study)

Save and close the workspace (Phuket ws), and from C:\Examples v7\Clyde macrofauna open Clyde environment, of 11 abiotic sediment variables (Cu, Mn, Co, Ni, Zn, Cd, Pb, Cr and %carbon and %nitrogen, plus water depth) sampled in 1983 at each of 12 sites (S1 to S12) along an E-W transect across the Garroch Head sludge dump-ground in the Firth of Clyde – see Fig. 1.5 of CiMC (data from Pearson TH, Blackstock J, 1984 Dunstaffnage Lab Report, Oban, Scotland). We will use these data (seen often in CiMC) for the rest of this section and most of the next one.

Missing data estimation

The subject of missing data has arisen several times already (Sections 1, 3, 5) and the point made that the terminology and sheet entry Missing! refers only to variables (usually environmental -type variables) that are not recorded for some samples. It does not refer to designs which were intended to be balanced but for which some replicate samples were not analysed for some reason, over all variables. (Unbalanced replication is not generally a problem to handle in PRIMER, since balance is not required for most of the testing that PRIMER, and PERMANOVA+, are able to carry out.)

Some of the routines, including PCA (next section), require the user to enter a complete matrix, with no missing values. At a simple level, it is fairly clear why this should be so. For the trivial 2-variable case in which PCA was introduced in Chapter 4 of CiMC, imagine losing one variable value for one of the samples. What is now that sample's contribution to total variance? How can it be projected perpendicularly to the best-fitting line through the points? How can that first PC axis be determined at all without knowing the contribution of this sample, and so on? In fact, a solution to this was suggested in Section 5 when discussing computation of resemblance measures in the presence of missing entries – it is possible to adjust Euclidean distances, or any other distance/dissimilarity measure, for the crude bias that may come (and certainly will come for Euclidean distance) from some pairs of samples having more matching variables across the two samples than others do. The resulting (near-)Euclidean resemblance matrix is then complete and a choice can be made between MDS (possibly metric) or PCO in the PERMANOVA+ add-on software. The latter is a PCA when the matrix is Euclidean (though the missing data will make that identity not quite true). An alternative is to remove (listwise) as few variables and samples as possible, in a judicious balance, such that a complete matrix is left. The routines **Tools>Check**, **Select>Samples>**(•No missing values) or **Select>Variables>**(•No missing values) will help with this. When there are large blocks of missing data – a subset of the variables were simply not recorded at a large group of sites – then this is likely to be the most realistic option. In other situations, where there is very little missing data, it can seem very wasteful of valuable resources – a whole sample would have to be deleted because one variable is missing, or a whole variable deleted because it was not measured for one sample. In this case, there are then two realistic options – work always from a resemblance matrix and allow PRIMER to adjust automatically the pairwise distances for the crude bias, or use a completed data matrix obtained by estimating the missing values with the EM algorithm. If some restrictive distributional assumptions apply (with rather few missing values and good correlations between some of the variables), this can provide a less crude adjustment and should be attempted.

v7

v7

EM algorithm assumptions

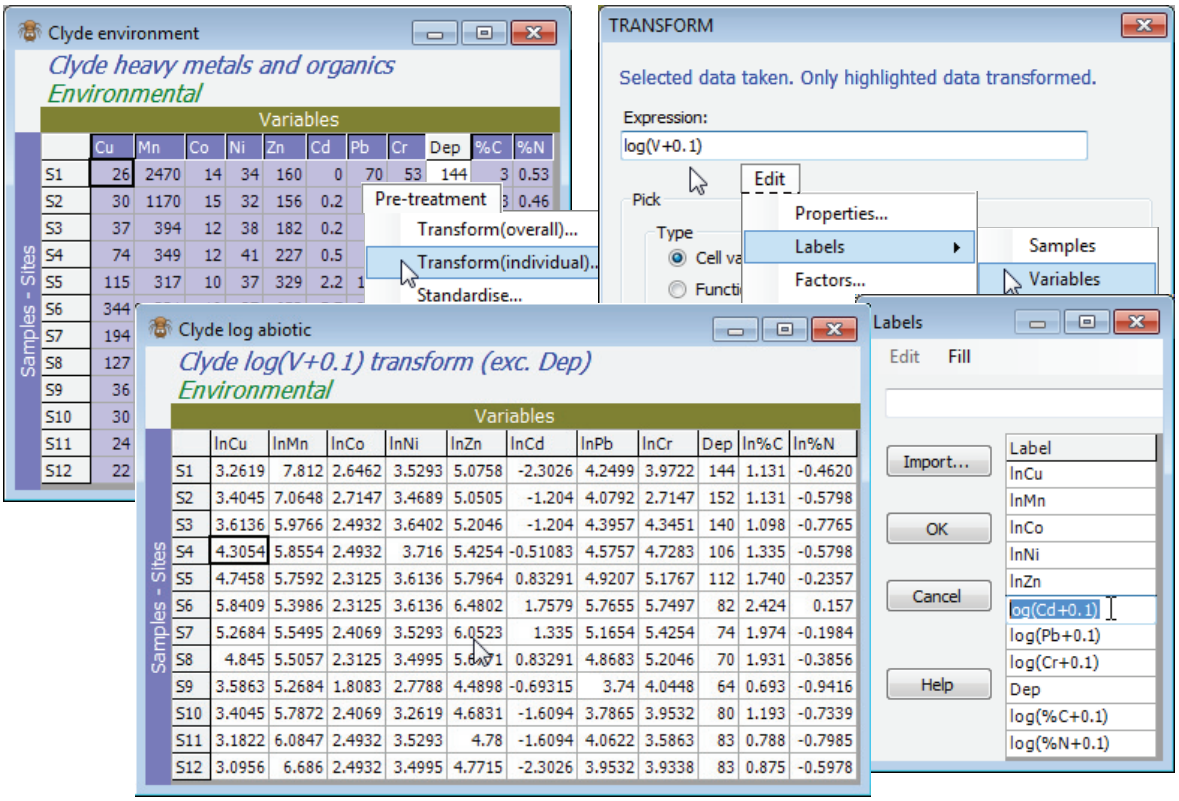
Tools>Missing is designed to operate only on matrices for which: a) assumptions of multivariate normality can be made; b) there are many fewer variables than samples, so that there are enough data values to be able to estimate the parameters representing means, variances and correlations of all the variables, with reasonable stability; c) there are rather few missing data points (each of those is a new parameter that needs estimating also); d) the data points are thought of as 'missing at random', rather than missing because they were so extreme that they could not be recorded; e) the samples are treated as of unstructured design, rather than, for example, utilising information about their status as replicates from a set of *a priori* defined groups.

Many of these are the assumptions that the methods of PRIMER are trying to get away from, of course! But that is mainly because they are completely impossible to satisfy for assemblage data; they may be much more realistic for continuous, environmental-type data (including, for example, morphometric variables). The estimation technique that PRIMER uses is the standard statistical method under these conditions, namely the EM (expectation-maximisation) algorithm. It is rather tricky (and dangerous!) to give guidelines for when the method will prove acceptable, but you do have some help from the algorithm. Firstly, if you set it an impossible problem (far too many parameters to estimate for the number of data points you have) then it will fail a convergence threshold and display an error message (*max number of iterations exceeded*). Secondly, when it does converge, it is also able to provide an approximate standard deviation for its estimate of each

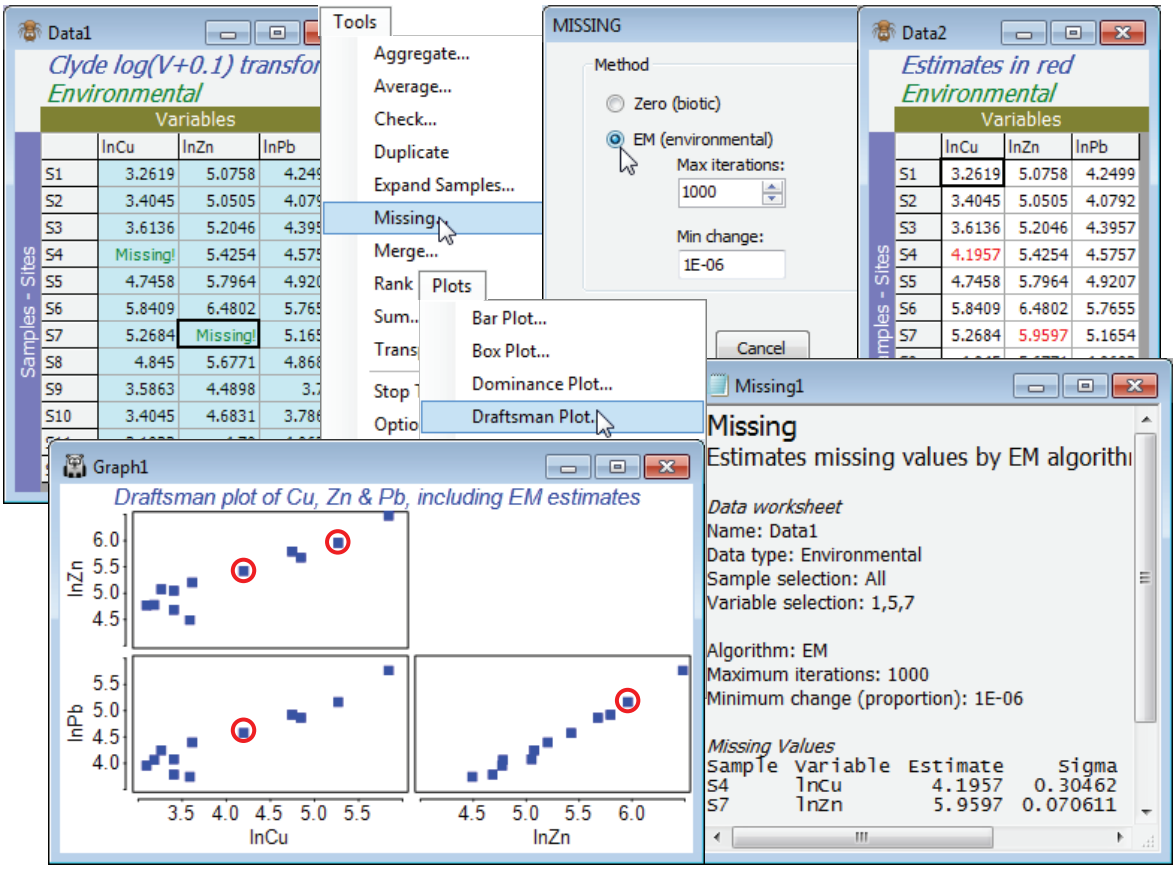
missing value. If this is large then there has clearly been insufficient information to pin down a likely value for the missing cell. As a rough rule-of-thumb, you should not expect to be estimating more than about 5% of your data points if your analysis is to retain any credibility(!), and you should have enough samples *n* compared with (selected) variables *p* and missing cells *m*, so that there is a half-decent number of data points per estimated parameter $DpP = n/[(p+3)/2 + (m/p)]$ (around 7 is sometimes cited, in general contexts). When this criterion is far from being met using the whole matrix, you may be able to take a piecemeal approach, selecting just a small set of the most relevant variables to drastically reduce *p*. The method is clearly only going to provide you with something useful if there are variables that correlate fairly well with the one containing the missing data, so that it has some basis for the prediction. **Draftsman Plot** will work on datasheets with missing cells, so you can use this (and its correlation table) to select out good subsets of variables for estimating each missing data cell. Use of **Tools>Missing** should not be seen as an automatic process therefore – you must expect to have to work hard to justify any data points that you are making up! In the end, common sense is the best guide here, as always. Look at each estimated value – they are always displayed in the worksheet in red – and compare it with the range of values from the other samples for that variable. Does it look 'reasonable', or has something clearly gone wrong with the fitting routine? If all appears well, then it does have the objective credibility of being the maximum likelihood estimate of that cell, and not just some subjective value that you wish it was! Also, look at the standard deviation (σ) of the estimate in the results window and try sensitivity analysis. Add or subtract up to 2σ from each of the estimated cell values at random, and re-run your PCA (or MDS, ANOSIM etc). Whatever you estimate for the missing values may make no difference to the outcome, if they are within a reasonable range of the other data – you then have a very credible analysis.

Missing data estimation (Clyde study)

Transformation options for the Clyde environmental matrix, Clyde environment, are discussed in more detail in the following (PCA) section, but the tool to carry out separate transforms on sets of variables, **Pre-treatment>Transform(individual)**, rather than transforming the whole array, **Pre-treatment>Transform(overall)**, was met in Section 4, applied to the environmental data from the Ekofisk oil-field study. Here, all heavy metals and organics (10 of the 11 variables) will benefit from log transformation, to reduce their right-skewness and so bring these continuous variables closer to normality across the sites (in so far as that can be judged from only 12 samples!). Thus, highlight all variables except Water Depth (*Dep*) and take **Pre-treatment>Transform(individual)** >(Expression: log(V+0.1)) & (✓Rename variables), renaming the result Clyde log abiotic. Give the variables in this sheet shorter names (e.g. lnCu, lnMn etc) with **Edit>Labels>Variables**.



Take a copy with **Tools>Duplicate** and from this remove a couple of cells at random – perhaps (S4, lnCu) and hit the delete key, then (S7, lnPb) and delete again. Both cells will now be displayed as Missing!. **Analyse>Draftsman Plot** on this transformed data shows that normality assumptions are probably now acceptable (see the following section) but the above *DpP* criterion for the whole matrix fails badly (n = 12, p = 11, m = 2, so *DpP* = 1.7) and we should not trust the outcome even if **Tools>Missing** converges (it does not, here). The correlation matrix output with the draftsman plot does, however, show some very high correlations between e.g. Cu, Pb and Zn, which gives a better basis for prediction than the whole matrix. So, select just these three variables (highlight them then **Select>Highlighted**), and **Tools>Missing** produces credible missing data estimates of 4.18 (S4, ln Cu) and 5.26 (S7, ln Pb), compared with the original 4.31 and 5.17. Note that the ratio *DpP* = 3.3, which is still some way from respectability, but clearly is capable (sometimes at least) of producing useful results. The results window shows that the imprecision (under the assumption that the value is missing at random, of course) is lower for the estimated (S7, Pb) reading than the (S4, Cu) value, though both are rather well determined. The standard deviation of the estimate for (S7, Pb) is about 0.07 and for (S4, Cu) about 0.30, so that rough confidence intervals are (3.6, 4.8) and (5.8, 6.1) respectively. The reason for this difference in precision is clear from the draftsman plot for these three Cu, Zn, Pb variables, on which the respective points are manually circled (the plot window was copied and pasted to Powerpoint with Ctrl-C and Ctrl-V). The linear relationship between Pb and one of the other variables (Pb) is seen to be extremely tight, whereas Cu is not so highly correlated with either Zn or Pb, so there is inevitably greater uncertainty in the interpolation – it is a consequence of the multivariate normality condition that these relationships are estimated as straight lines. The estimates now need to be individually copied (click in the cell and Ctrl-C) and pasted back into the full matrix (Ctrl-V at the cursor). Of course the process is more automatic in less borderline cases, with larger n, when the full matrix can be input to **Tools>Missing**.



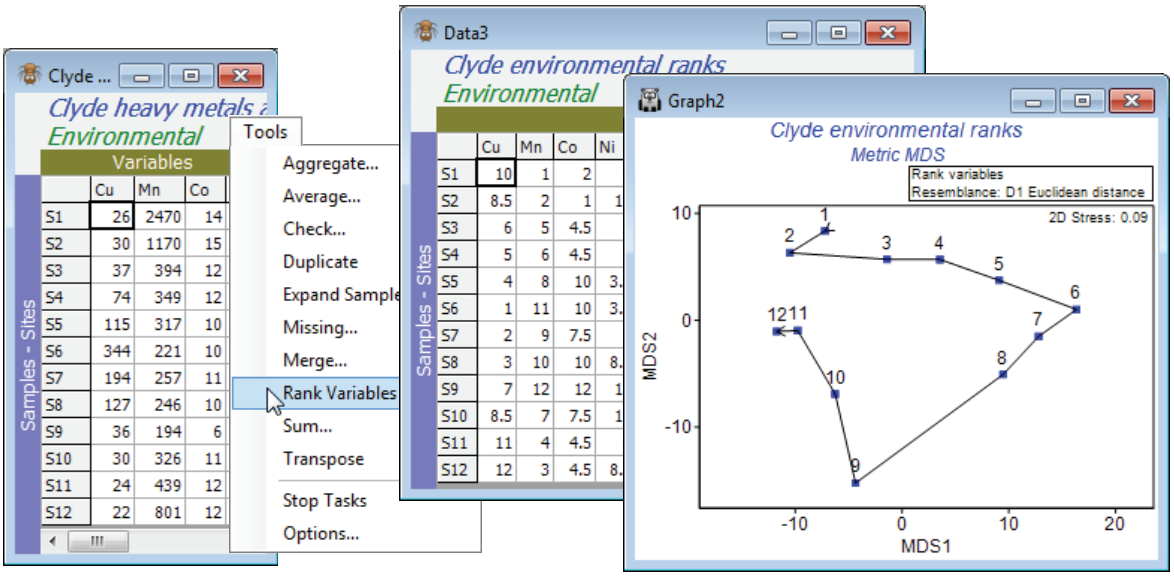
Ranked variables

The following section (on PCA) will discuss further the choice of particular transformations to avoid the sensitivity of PCA (and Euclidean distances in general) to outliers in some environmental variables, but choice of individual transformations is often a worry to practitioners. An alternative, eliminating the need for choice (but arguably losing some sensitivity in the ensuing analysis), is to replace variables by their ranks, namely the numbers 1, 2, 3, … for largest to smallest values across samples (modified if necessary to substitute average ranks for tied values). The main advantage is

that the over-dominant contribution of outliers is automatically eliminated. For example, a variable whose values over the samples, in decreasing order, are: 25, 9, 7, 6, 6, 6, 4, 2, 2, 0 would generate ranks: 1, 2, 3, 5, 5, 5, 7, 8.5, 8.5, 10 respectively, and the effect is to make the outlying value of 25 no different than if it had been 15 or 10. Ranking each variable (separately) also removes the need for normalising the resulting array, which is needed (after transformation) with the usual approach, to ensure that all environmental variables take values across comparable ranges. Ranking places all variables on a common measurement scale, the numbers 1 to *n* (where *n* is the number of samples).

For the original (complete) Clyde environment sheet, take **Tools>Rank variables** and examine the outcome. Put this matrix through **Analyse>Resemblance>**(Measure•Euclidean distance) and then **Analyse>MDS** for a non-metric or metric MDS (the latter has a better chance of being acceptable because of the few points and the simple gradient structure, and importantly, the Euclidean distance matrix). In order to overlay a trajectory on the MDS with **Graph>Special>Overlays>**(✓Overlay trajectory)>(Trajectory numeric factor: Site#), you will need either to create the *Site#* factor for any sheet on the Clyde environment branch, with **Edit>Factors>Add>**(Add factor name: Site#), highlighting the column and **Fill>Label number**, to generate the values 1 to 12. (Alternatively, if you have already opened the abundance file Clyde macrofauna counts into the workspace, you can **Factors>Import** the factor *Site#* from that sheet). It is interesting to note the linearity of Shepard diagrams for both *m*MDS and *n*MDS but whilst the ordinations look very similar, the *m*MDS fit of a straight line through the origin is not quite such a good fit (stress = 0.09 c.f. *n*MDS stress = 0.03). The main point here, though, is that this ordination, based on ranked data, looks very similar to the PCA which we shall see in Section 12, based on transformation and normalisation of this data.

v7

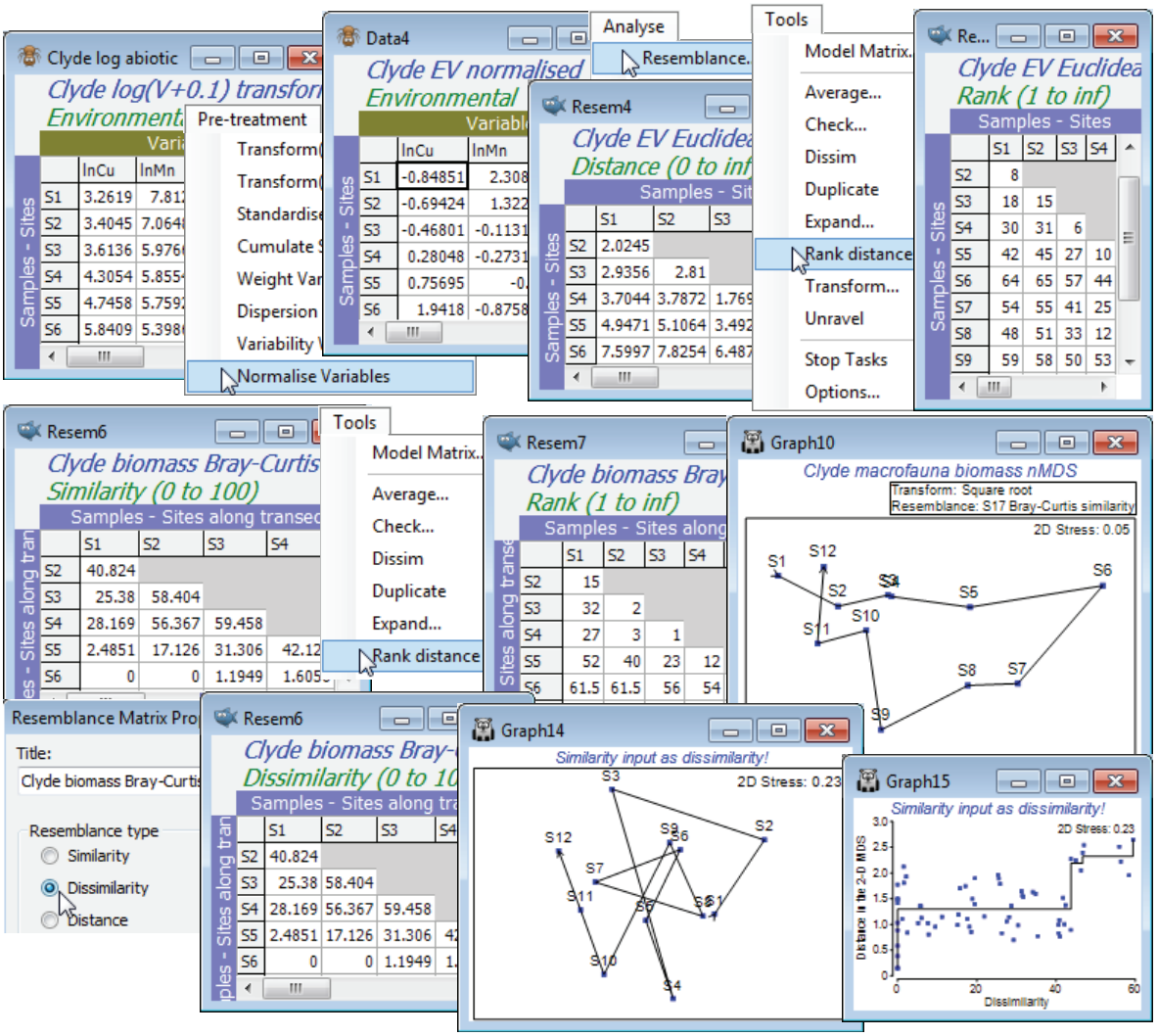


Ranked resemblances

Ranking is also a menu option when the active sheet is a resemblance (**Tools>Rank distance**), but it operates a little differently. This time, all elements of the triangular matrix are ranked together, rather than separate ranking of the rows or columns of the rectangular data sheet. Do not get these two possible rankings confused! It is easy to fall into the trap of thinking that, because a ranked data matrix will be the same whether ranked from original or transformed data, if you are intending to rank the similarity matrix then initial transformation of the data does not matter. This is entirely wrong of course – ranking the similarity matrix is by no means the same as ranking the data then calculating the similarities! In fact, whilst ranking the data may play a marginally useful role for handling outliers in environmental matrices (as above), it rarely makes sense for assemblage data because it destroys the special nature of the (very many) zero responses, which would be assigned different tied ranks for different species. Ranking the resemblances, however, is rather central to the approach in PRIMER: many of the core routines (ANOSIM, RELATE, BEST, …) start from the ranked form of the similarity matrix, and *n*MDS ordination also exploits this rank order. For all routines, however, it is not necessary to enter the ranked form of the triangular matrix – if the result depends only on the ranks, this will be part of the internal calculation on the similarities. The menu item of **Tools>Rank distance** on a resemblance matrix is mainly here to help visualise and check the relatively simple computations underlying an ANOSIM test, for example (see the definition of the ANOSIM R statistic, a difference in mean rank dissimilarities, in equation 6.1 of CiMC).

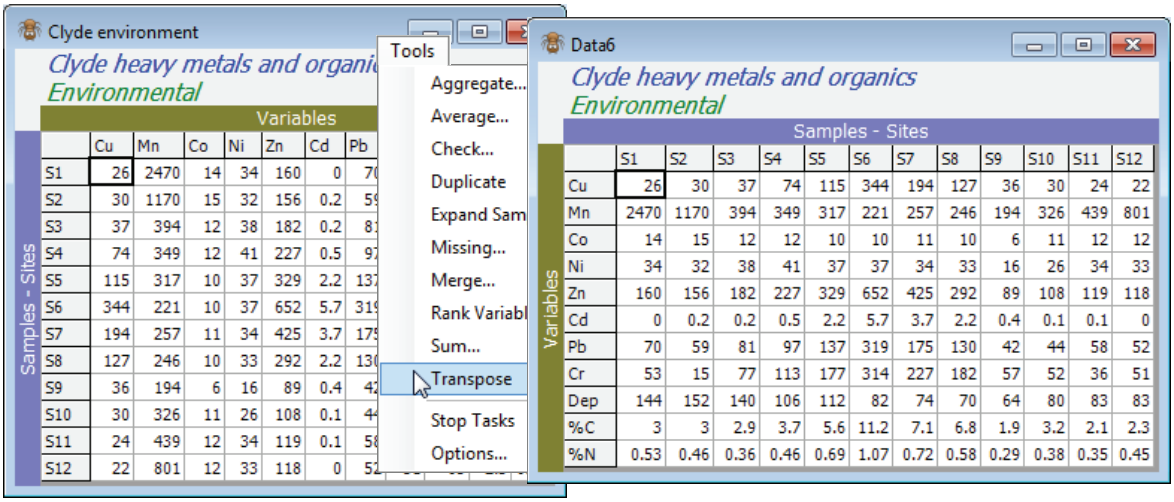
For the Euclidean distance matrix from the above Clyde environmental data (transformed then normalised), take **Tools>Rank distance** to produce a rank resemblance matrix. Note that entries are just the numbers 1, 2, .., 66. Importantly, the convention PRIMER adopts here is always to return a distance-type matrix from the **Rank distance** operation, irrespective of whether it is given a similarity or dissimilarity/distance matrix (explaining why the menu item is called **Tools>Rank distance**). Thus rank 1 corresponds to samples (S11, S12), which are closest environmentally, and rank 66 to those furthest apart (S6, S9). To see this point about the direction in which ranks are assigned, open the macrofaunal biomass matrix for the 12 samples on the Clyde transect, Clyde macrofauna biomass, take a square transform and calculate Bray-Curtis similarity, as usual. Now take **Tools>Rank distance** on this similarity matrix and note that the resulting ranks again form a distance matrix, with the closest sites in assemblage terms (rank 1) being S3 and S4, and several pairs of sites tied on the largest, most distant rank (average of 61.5), namely S6 with S1, S2, S11, S12 etc, which are all pairs of sites with no species in common.

PRIMER handles its (distance) ranks in this slightly unconventional way to reassure the user that, on the many occasions when two sets of resemblances are compared to see if they are arranging the samples in a similar high-dimensional pattern (e.g. assemblage *vs* environment, Bray-Curtis *vs* Chi-squared or Euclidean measures, biomarkers *vs* tissue burdens etc), the user does not have to worry whether the two resemblance matrices are the 'same way round' (whether high values correspond to large or small differences between samples). This is always adjusted to the correct comparison, in the same way that the MDS routine will always internally turn a similarity into a dissimilarity when it is matching this up to distance in the ordination space (as in the Shepard diagram). You can force PRIMER to do the stupid thing, e.g. run MDS the 'wrong way round', making it try to place sites that should be similar at the greatest distance apart and sites that have little in common close together (with resultant very high stress levels, and a crazy plot and Shepard diagram!). But you can only do this by giving PRIMER a genuine similarity matrix and calling it a dissimilarity, by using **Edit>Properties** to change (Resemblance type•Similarity) to (•Dissimilarity).



Transposing the datasheet

The Clyde environment sheet has samples as rows and variables as columns. This is the opposite of the ecological matrices typically seen so far, such as Clyde macrofauna biomass, in which rows are the variables (species). The environment matrix is displayed according to the convention in classic multivariate statistics (samples as rows) but ecologists, for good reason, have long chosen to use the transposed form. This is because they often have p (species) > n (samples), whereas classical (normality-based) multivariate methods require $n >> p$, and it is generally neater to put the larger set of labels into rows (this also suits lengthy species names). It makes no difference to PRIMER which way round the matrices are held, the only important specification being which axis holds the samples (rows or columns?). That is changed by (Samples as•Columns) or (Samples as•Rows) on the **Edit>Properties** menu and not by transposing the array (so that columns turn to rows and rows to columns). However, a **Tools>Transpose** operation may sometimes be helpful in displaying a sheet in PRIMER or, more likely, before saving the data to an external file, when another software application needs a particular orientation. Take **Tools>Transpose** on Clyde environment and note that the Samples/Variables designation also switches.



Transform (individual) advanced

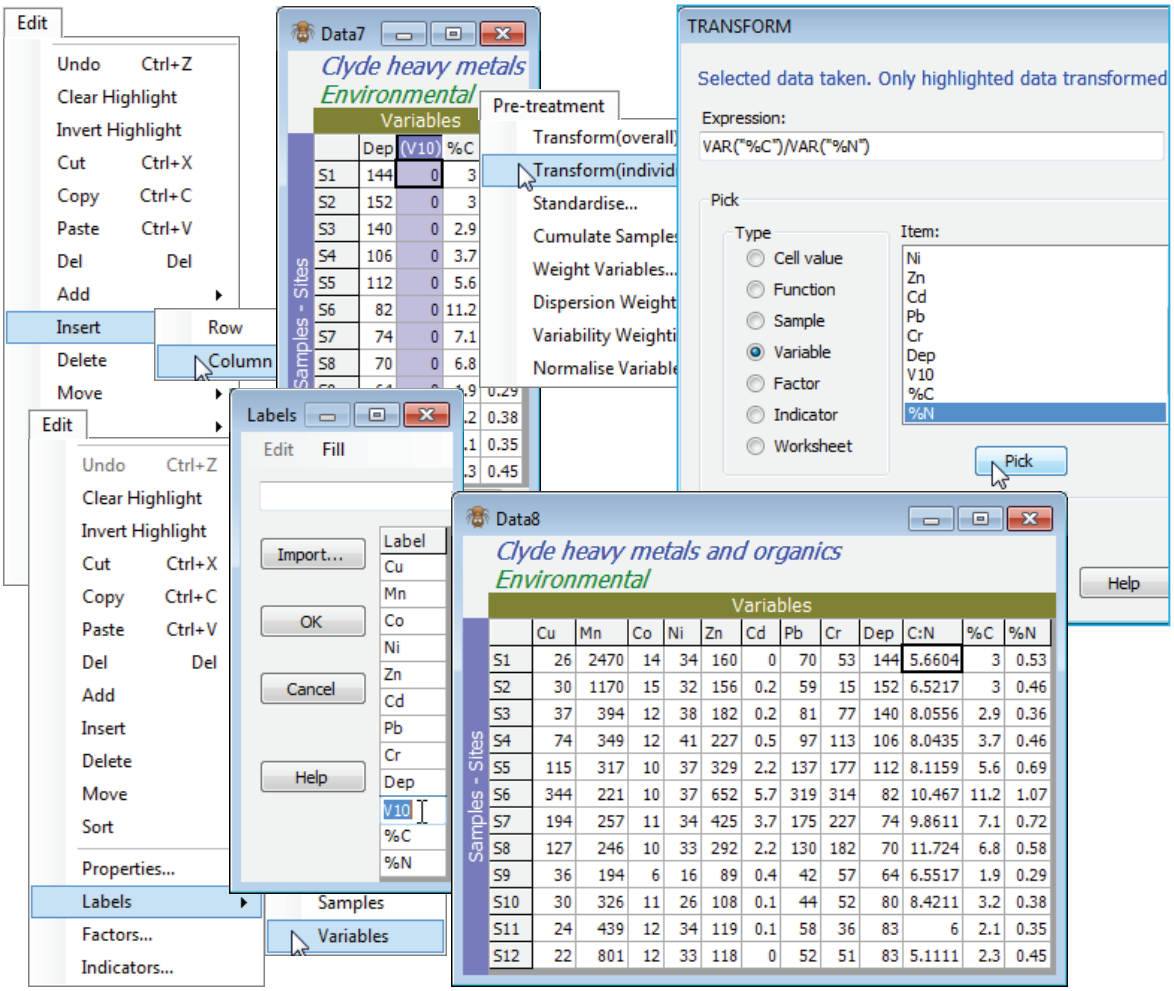
Unlike previous versions, in PRIMER 7 the **Transform(individual)** routine has been moved to a more convenient – and logical – position in the **Pre-treatment** menu. Its routine use is therefore covered in Section 4, and its application has been seen several times already. However, in order not to break up the presentational flow of a typical analysis pathway in this earlier section, the more complex features of this routine were deferred to this section. As a brief recap, **Pre-treatment> Transform(individual)** operates on highlighted, not selected, portions of the data sheet (if there is no highlighting it takes place on the entire sheet) and produces a new sheet according to a BASIC language-type (Expression:) provided by the user, in which V stands for the existing value in each cell which is being operated upon. A Pick>Type list aids in the construction of expressions by providing a suite of possible functions (•Function), some of which are standard BASIC definitions (LOG(V), EXP(V), INT(V), ... – note that the difference between upper or lower case is ignored) and some are designed specifically for commonly-used operations (e.g. ARCSINE(V) is the often seen arcsin transformation – more often seen than is justified in fact! – in which the exponent is first square-rooted before arcsin, the ASIN(V) function, is applied; these are new to PRIMER 7). The Pick>Type list also has the facility to use the values of an existing (•Sample), (•Variable), (•Factor), (•Indicator) or even whole (•Worksheet), so there is much flexibility to manipulate a data matrix to a new form, totally within PRIMER. Having said that, many users will still find it more convenient for very complex operations to use the tools they are already familiar with outside the package – e.g. in Excel – but saving data to Excel, manipulating and re-opening it in PRIMER is a relatively painless procedure, since Excel moved away from its 255 column limit! (PRIMER v6 and beyond do not have any fixed restrictions on data sheet sizes but are inevitably limited by the available RAM and by execution time for routines such as MDS and SIMPROF, as noted earlier).

v7

Expressions combining variables

For an example of an Expression combining two (or more) variables, use the Clyde environmental sheet but copy it (**Tools>Duplicate**), which is always a good idea when experimenting! The aim is to create a new variable (column) which is the C:N ratio, so first **Edit>Insert>Column**, which will be placed to the left of the current cursor position – here this might logically be on the %C column.

[Incidentally, remember that the new **Edit>Undo** will step back any **Edit** operations like this which change the current data matrix, rather than creating a new sheet (where a new sheet is created it can always be deleted if incorrect, and the process repeated). And note that **Edit>Undo** is local to the currently active matrix, so will undo the last such operation on this sheet, irrespective of whether similar operations have been performed since, on other sheets in the workspace – they will not be
v7 wound back by **Edit>Undo** if their sheet is not the active one]. The new column is labelled (V10) because of its position in the matrix, and the calculation we do to create the C:N ratio is to be put in this column, so it needs to be highlighted. Take **Pre-treatment>Transform(individual)** and delete the V from the Expression box (that refers only to values in the new column, which we shall not be using – they are all zero of course). Then, under Pick, take (Type•Variable) & (Item: %C)>**Pick**) and follow by (Type•Variable) & (Item: %N)>**Pick**), which creates the two variables we need in the Expression box. Manually insert the divide symbol (/) between them, to give the (Expression: VAR("%C")/VAR("%N")), and **OK** now gives a new sheet with the added C:N ratio variable V10. (Even if you chose to tick the (✓Rename variables) box, the new name will still be clumsy and it would be better to change it to C:N using **Edit>Labels>Variables**).



	Cu	Mn	Co	Ni	Zn	Cd	Pb	Cr	Dep	C:N	%C	%N
S1	26	2470	14	34	160	0	70	53	144	5.6604	3	0.53
S2	30	1170	15	32	156	0.2	59	15	152	6.5217	3	0.46
S3	37	394	12	38	182	0.2	81	77	140	8.0556	2.9	0.36
S4	74	349	12	41	227	0.5	97	113	106	8.0435	3.7	0.46
S5	115	317	10	37	329	2.2	137	177	112	8.1159	5.6	0.69
S6	344	221	10	37	652	5.7	319	314	82	10.467	11.2	1.07
S7	194	257	11	34	425	3.7	175	227	74	9.8611	7.1	0.72
S8	127	246	10	33	292	2.2	130	182	70	11.724	6.8	0.58
S9	36	194	6	16	89	0.4	42	57	64	6.5517	1.9	0.29
S10	30	326	11	26	108	0.1	44	52	80	8.4211	3.2	0.38
S11	24	439	12	34	119	0.1	58	36	83	6	2.1	0.35
S12	22	801	12	33	118	0	52	51	83	5.1111	2.3	0.45

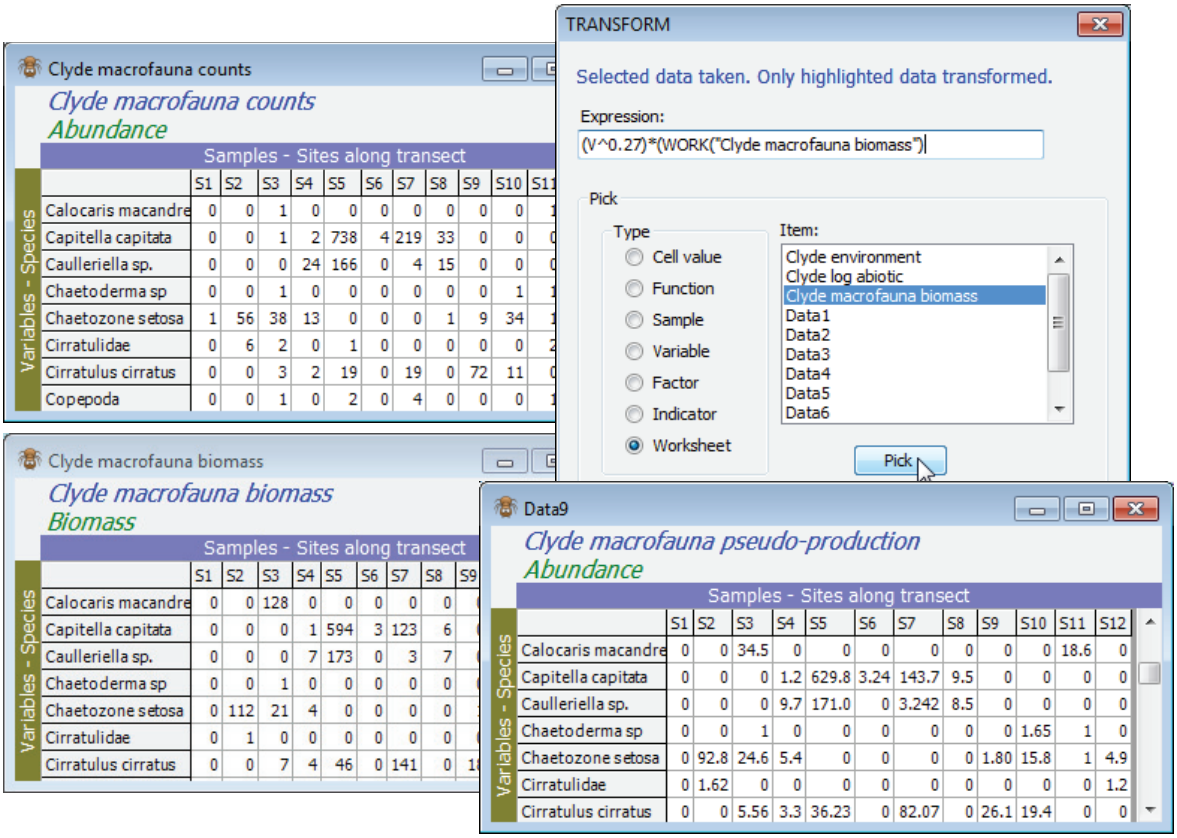
An alternative, e.g. if you just intend to take this new variable back into the earlier transformed sheet, is not to insert a new blank column, instead just highlighting the %C column (which will now be V in the Expression box), and **Pre-treatment>Transform(individual)** with (Expression: V/VAR("%N")). In the new sheet, this will have overwritten the %C row with the C:N ratio. Either way, you can now put the new C:N variable back into the transformed sheet simply by highlighting it, then **Select>Highlighted** and **Tools>Merge** it with the transformed sheet, taking the defaults.

Expressions combining worksheets

Similarly, expressions can combine samples, or even factors (or indicators) on those samples (or variables) – and expressions can even incorporate different worksheets. In fact some of the most useful applications of complex expressions are in combinations of data from related worksheets, such as the abundance and biomass arrays of macrofaunal assemblages from the Clyde study. The key facts to keep in mind when constructing complex expressions are that V stands for any entry in

the active sheet that is highlighted, that the result will be placed only into these highlighted cells (which could mean the whole array, if there is no highlighting), and that maintaining strict labelling across worksheets will make it easier to understand what the expression calculates for each cell. (Though, as elsewhere in PRIMER, if **Transform(individual)** is given two data sheets that have conformable dimensions but not consistent labelling, then it will give the user the chance to relax strict matching and assume that samples or variables are presented in the same order).

If not already in the workspace, open Clyde macrofauna counts and Clyde macrofauna biomass. One useful way of combining abundance and biomass information from the same set of samples is in equation (15.1) of CiMC, namely an allometric equation for *pseudo-production*, $P = A^{0.27}B^{0.73}$. With the abundance sheet active, turn off any highlighting with **Edit>Clear Highlight**, or highlight everything (the effect is the same), then take **Pre-treatment>Transform (individual)**>Expression: (V^0.27)*(Work("Clyde macrofauna biomass")^0.73). You can use (Type•Worksheet) & (Item: Clyde macrofauna biomass)>**Pick** to give you the syntax for the second sheet but type the rest; the counts of the first sheet are held in V (≡ Work("Clyde macrofauna counts")), since that is active. You should definitely uncheck (✓Rename variables) so that the species names remain intact.

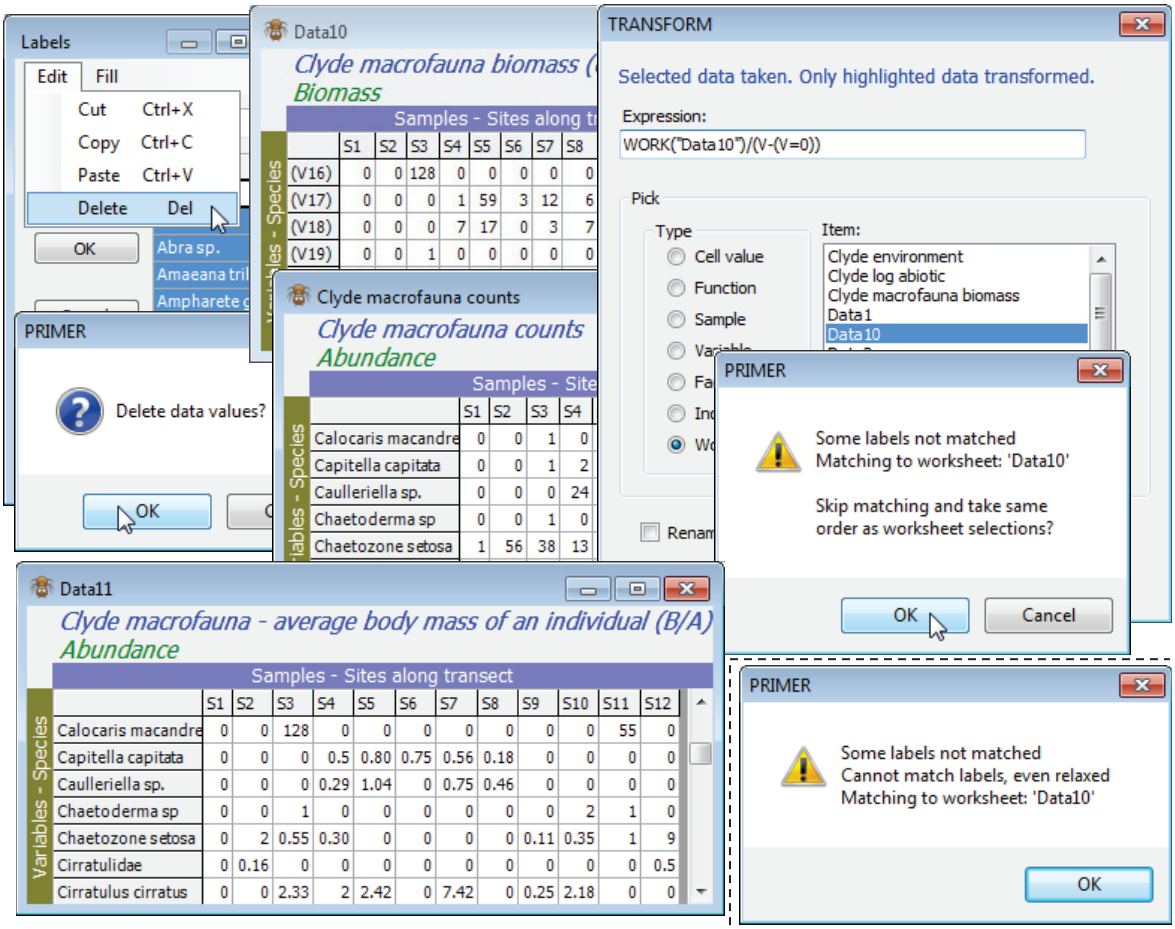


Average body mass matrix (B/A)

A useful variation of this, but one which needs more care, is to compute average body mass of each species in each sample. This is simply *B*/*A*, but needs to cater for the many cases when *A* (and *B*) are zero and a simple ratio is undefined. With active sheet Clyde macrofauna counts, so that V is again the counts, **Pre-treatment>Transform(individual)**>Expression: Work("Clyde macrofauna biomass")/(V – (V=0)) will do the trick, because when V>0 the expression (V=0) gives the value 0 (false), so that the correct ratio of *B*/*A* is calculated. However, when V=0 the expression (V=0) returns the value -1 (true). The bottom line is then 1 and the result of the ratio is a reasonable value of 0. This assumes that *B*=0 when *A*=0 of course! [This, incidentally, is something that can be checked by running Abundance-Biomass Comparison curves, described in Section 16, since the **Analyse>Dominance Plot** (ABC) routine explicitly checks for incorrect matrix entries which have *A*=0 but *B*>0; the converse is perfectly permissible – the weight of all organisms of a species in a sample might be too small to register – but this does not cause a problem with a *B*/*A* calculation).]

An illustration of error trapping and relaxation of strict matching, in **Pre-treatment>Transform (individual)** with matching of entries, is obtained by copying Clyde macrofauna biomass with **Tools>Duplicate**, then **Edit>Labels>Variables** on this to delete all the species labels (click the

Label header and hit the delete key or **Edit>Delete**). A sheet cannot function without labels so PRIMER substitutes its own defaults of (V1), (V2), etc. Now run the above calculation on Clyde macrofauna counts, but with the relabelled biomass sheet (Data10 below) replacing the original biomass sheet. A warning message says that it could not find (variable) labels to match, but the two matrices are the same size so the option is given of proceeding anyway, on the assumption that the species order matches. We know it does here, so continue, to give the desired *B*/*A* matrix, and the original species labels will be present in the resulting new sheet because these are always taken from the active matrix, in a case such as this. Re-run having deselected one of the rows in Data10, however, and an irrecoverable error message occurs – a match is impossible because the variable labels do not match and neither does the number of variables in the *A* and *B* matrices.



Transform on resemblances

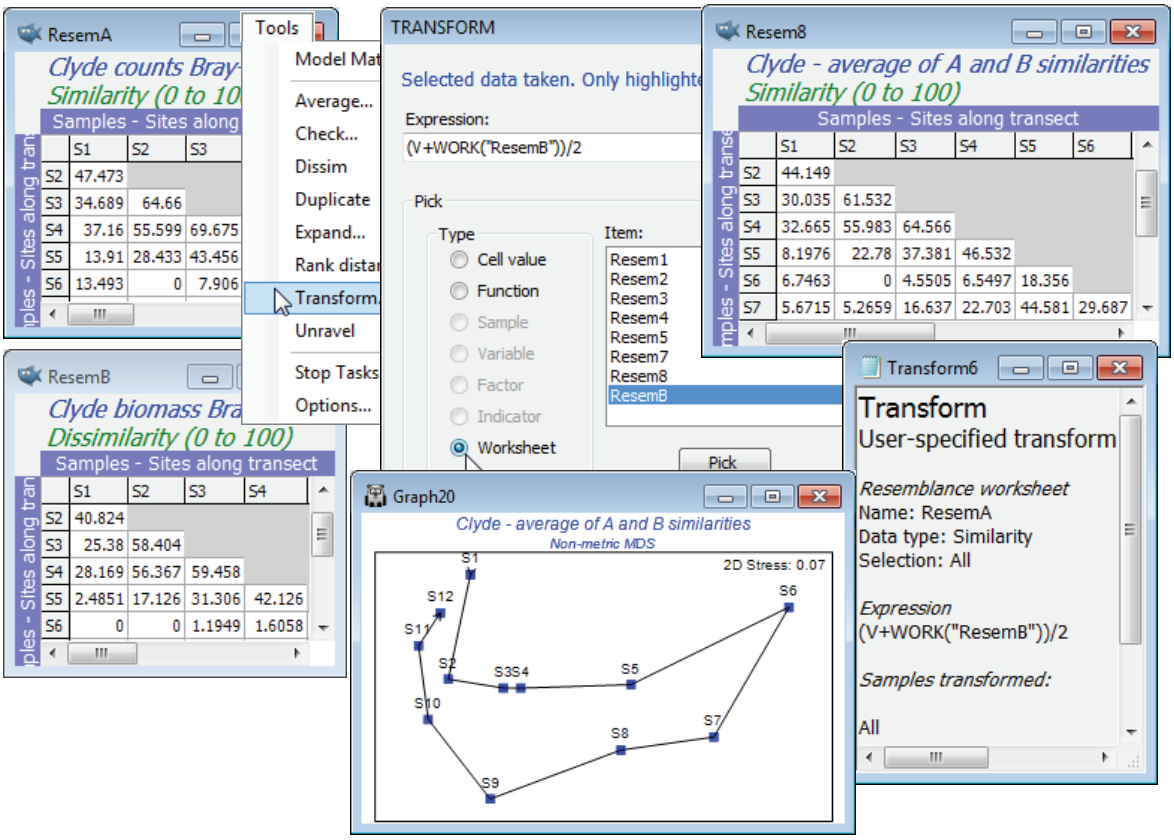
Transforming resemblances remains in the **Tools** menu in PRIMER 7, since it is not an option for pre-treatment of data matrices prior to resemblance calculation (which characterises the other items on the **Pre-treatment** menu). Although not commonly required, it facilitates at least a couple of interesting analysis concepts. One is really outside the scope of this manual, namely to examine the extent to which the semi-parametric PERMANOVA tests are robust to (monotonic) transformation of the resemblance values, transformations which would not change the ANOSIM test results in any way (since they are based only on the ranks of the resemblances). It is empirically well-known, for example, that the square root of Bray-Curtis (unlike Bray-Curtis itself) does not give negative eigenvalues for the high-d PCO ordination which underpins the approaches in PERMANOVA+. Whether the consequentially poorer low-d PCO representation is a price worth paying for a PCO space without imaginary axes must be open to question, however. **Tools>Transform** on a single resemblance matrix provides a basic tool to assist in following up such questions. More simply, we have already seen it used (under Correlation as similarity in Section 5) to turn a correlation with values in (-1, 1) into a similarity over (0, 100) by use of the transform Expression: 100*Abs(V).

Combining resemblances

Another use of **Tools>Transform** on resemblance matrices is also less esoteric and potentially of substantial practical benefit. It provides an interesting solution to the handling of ecological data matrices from mixed faunal types, e.g. counts of motile organisms and cover of colonial species within the same rocky-shore quadrats. This type of problem was raised earlier (end of Section 8),

when two resemblance matrices over the same set of samples were combined in a single MDS, by minimising an average stress function. The difficulty with that approach is that it only generates an MDS, and many of the methods in PRIMER do not work in the approximate low-d space of an ordination but on the full resemblance matrix (or usually its ranks). However, whilst counts and area covers are difficult to scale in relation to each other in a single data matrix, it is not difficult to calculate Bray-Curtis similarities (say) for a count matrix and a cover matrix separately, for the same set of samples, and then simply average the two resemblance matrices over every matching pair of entries using **Tools>Transform**, using a similar worksheet-based transform expression to that previously demonstrated. (Dis)similarity values in the range (0, 100) for both matrices will stay in (0, 100) under the arithmetic averaging expression of $(A+B)/2$ (or a weighted form, $(3*A+B)/4$, if the contribution of counts is considered roughly three times as important as that from area cover). Geometric averaging of the type seen above is also possible, e.g. $(A*B)^{0.5}$ or $(A^{0.75})*(B^{0.25})$. If the two resemblance matrices are not on a common scale and direct averaging is not appropriate, a simple solution would be to run both through **Tools>Rank distance**, putting them on a common scale – and fitting well with PRIMER's non-parametric approach – then averaging as above and re-ranking (though the latter is unnecessary for most PRIMER routines, which do their own ranking).

A simple example using two resemblance matrices can be constructed with the Clyde data, namely the Bray-Curtis site similarities averaged over abundance and biomass measures. So, instead of combining the data matrices (as in the earlier $A^{0.25}B^{0.75}$), we average the A and B resemblances. There is likely already to be a Bray-Curtis similarity matrix based on the square-root transformed biomass data from Clyde macrofauna biomass in the workspace (rename it ResemB) and you should now also compute Bray-Curtis similarities from Clyde macrofauna counts, this time on fourth-root transformed data (there is no reason why a different transform should not be appropriate for abundance than for biomass data). Then with the abundance similarities ResemA as the active sheet, take **Tools>Transform>**(Expression: (V+Work("ResemB"))/2), change the title of the result appropriately, and run **Analyse>MDS** to compare this with the earlier (ranked) biomass MDS.



Tools menu - other items

v7

Tools operations on resemblances which are discussed elsewhere are: a) **Dissim** and **Unravel** in Sections 5 & 6 – the former turns similarity into dissimilarity, or vice-versa, and the latter creates a single column of entries from unravelling rows of the triangular matrix; b) **Model Matrix** and **Expand** in Section 14 – these are less trivial and need contextual explanation; and c) **Stop Tasks** in Section 6 (self-explanatory). That leaves only **Tools>Options**, PRIMER's default settings.

Tools Options menu

Options appears on the **Tools** menu whatever the type of active window. The items on the **File** tab were seen in Section 1, namely the default setting for the initial directory on launch of PRIMER from the Windows desktop (in general there is not a strong incentive to change this from its default blank entry since, once closed, PRIMER will relaunch to the last used directory), and whether, on opening workspaces, they are displayed with full branches unfurled or not (this can take some time for very large workspaces). There is also the option to suppress the initial dialog about this feature, which is otherwise displayed every time a workspace is opened. When on this, or any of the other tabs, the *factory default* settings for PRIMER (for all tabs) can be re-instated by **Reset defaults**, and those defaults are as illustrated in the dialog boxes below.

v7

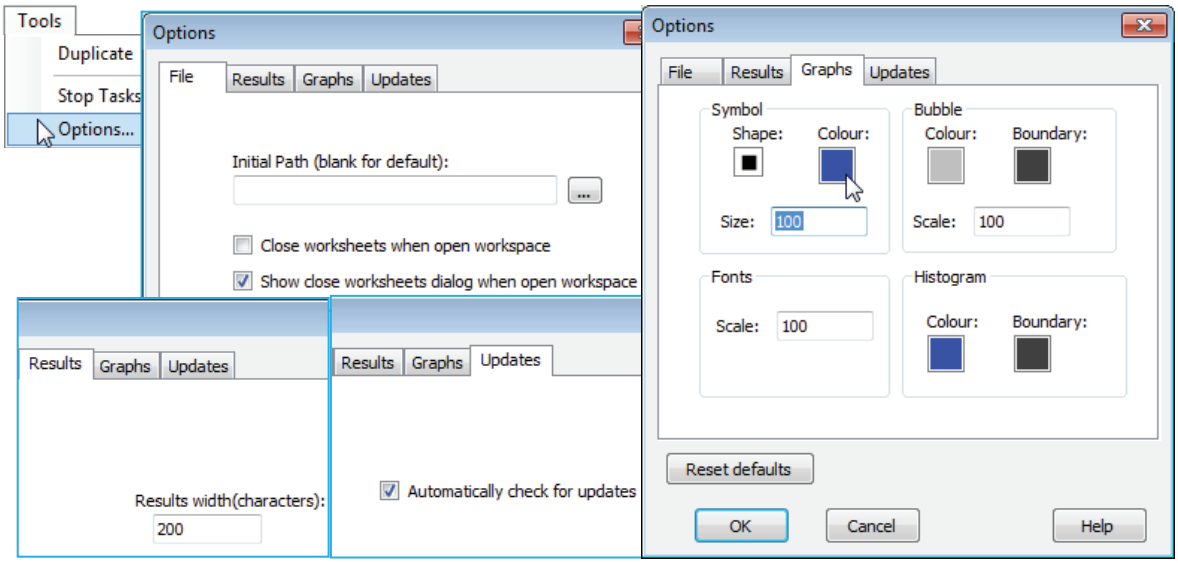
The **Results** tab contains a single, little-used item, which just determines the page width for results – the number of characters in the fixed-spaced font used for Results windows. This is initially set to 200 and only comes into play with the few routines (SIMPER and DIVERSE, Sections 10 and 15) which can generate wide lists of results. If this default is set to a smaller value than will allow a single span of results columns, they are split and listed a batch at a time. In practice, the DIVERSE routine essentially produces a matrix of samples (rows) by diversity indices (columns), so it will usually be preferable to direct this to a new worksheet, where it can be further plotted or analysed as a multivariate array (or be exported to Excel or a univariate statistics package). The **Updates** tab similarly concerns a single issue, this time a check box (new to PRIMER 7) to specify whether the software should automatically check the PRIMER-e server for existence of a maintenance update, and if it finds one, to ask the user whether they wish to download this at that point in time or not.

v7

The **Graphs** tab is the most likely to be used on a regular basis, since this sets some of the global defaults for all plots. In the Symbol area, Shape, Colour and Size can be set for all graphs on which a single symbol type is plotted, e.g. a draftsman plot, Shepard diagram, an ordination graph which does not plot different symbol types by factor or, indeed, a bubble plot for a single variable with no factor used (with a factor, or for >1 variable, the bubble colours are determined by the appropriate Key dialog). The Bubble area therefore only controls the inner and boundary colours for the bubble key when it accompanies a bubble plot utilising more than one different coloured bubble, hence the default of a neutral grey with black boundary to avoid a misleading colour synchrony with any of the factor levels. The Bubble Scale box does, however, apply to all bubbles because it sets the size for a bubble at the maximum variable value given by the Bubble key. In fact, it provides the default for the (Bubble scale:) box in the **Graph>Special>Main>**Bubble area, and the Symbol Size default similarly sets the value in the Symbol area of the **Graph>Sample Labels & Symbols** tab (hence applies also to symbol sizes when plotted by factor levels). The Fonts Scale box likewise fixes the displayed default for (Overall font scale:) on the **Graph>General** tab, so applies to all fonts. Lastly this **Graphs** tab sets the default Histogram inner and boundary colours. It must be appreciated that default changes are not retrospective; they apply only to plots created after any default changes.

v7

v7



12. Analysing environmental variables (*Draftsman Plot, PCA*)

Environment -type data

PRIMER uses the term *environmental variables* as a shorthand for a wide variety of data types (including biological data!), extending well beyond the archetypal case of physical or chemical measurements made on the environment surrounding an assemblage sample. Environment-type variables can also include matrices of biomarker responses (biochemical, sub-cellular or whole body health indicators from individual organisms, Section 4), morphometric measurements on individuals (perhaps with the aim of separating putative species), PSA data (size-class spectra for soil/sediment/water particulates, Section 4), organism body-size distributions, etc. The unifying factor for these disparate examples is that: a) they all give rise to multivariate arrays of variables by samples which can be analysed by the methods in PRIMER; b) the criteria which lead to use of community-type similarity measures such as Bray-Curtis are not appropriate (e.g. always positive entries, with many zeros and zero playing a special role – joint absences carrying no information, samples with no species in common having zero similarity – and always a common measurement scale across variables, of abundance, biomass, % cover etc). Instead, resemblance between samples of environment-type variables is better described by standard distance measures such as Euclidean distance (Section 5), where zero plays no special role (e.g. zero temperature, but on what scale?), where negative values can occur (indeed will occur if normalising different scales to common units, Section 4), and where positive similarity is always inferred if two samples have the same value of a variable, even a zero value (e.g. neither sample has a detectable PCB or Hg level, neither sample has particles > size x, etc). The key message here is that whole assemblage data is different, and requires the specialised methods that are at the core of PRIMER (biological similarity coefficients, non-metric MDS plots, non-parametric ANOSIM tests etc), environmental-type data is more standard and is often (after individual transformations and normalisation) best treated by the more classic approaches of Euclidean distances and Principal Components (PCA) ordination. The derivation and purpose of PCA is covered in detail in Chapter 4 of CiMC.

Draftsman plots recap & transform choices

Normalisation (subtracting the mean and dividing by the standard deviation, for each variable), and subsequent selection of Euclidean distance or PCA, operates more effectively the closer the data is to approximate (multivariate) normality. The latter is not a prerequisite of PCA but it is the genesis of the method and it is certainly true that, if the data is strongly skewed, the outliers will dominate the PC axes and will often lead to poor-quality interpretation. Transformations of specific variables, or groups of similar variables will often be desirable, by **Pre-treatment>Transform(individual)** – as in the previous section, and first met for environmental variables in Section 4. A useful aid to transformation choice is given by **Plots>Histogram Plot** or, where there are fewer samples, **Plots> Draftsman Plot**. The latter gives pairwise scatter plots between all (selected) variables. Two things are being looked out for here. Firstly, in the draftsman plot, are the samples roughly symmetrically distributed across the range of each variable? Or, if there is enough data to plot sensible histograms, are they very roughly bell-shaped, or at least symmetric rather than strongly skewed to one side? Secondly, if there are strong relationships between some pairs of variables, are these roughly linear rather than strongly curvilinear? This is also characteristic of (approximate) multivariate normality and an underpinning assumption of PCA, that ordinary product-moment correlations describe the dependence between variables (standard correlation measures only linear relationship). Examining these plots can therefore suggest possible transformations. If a distribution is right-skewed (bulk of the distribution to the left, with stragglers to the right) then a $\sqrt{y}$ (mild) or $\log y$ (strong) transform is called for. Use $\log(c+y)$ if y can be zero or negative, choosing a constant c to make all the $(c+y)$ values strictly positive before taking the log. If it is heavily skewed to the left, consider an inverse transform, $1/(c+y)$ where c is close to zero, or a reverse transform, $\log(c–y)$ or $\sqrt{(c–y)}$ (strong or milder), where c is chosen to be larger than the maximum y. Try to use similar transforms for the same types of variables, and don't be too pernickety! Logically, you need to use the same transform each time you analyse new data in the same context, and over-detailed choices will preclude that. The idea is only to avoid the worst effects of extreme outliers when working on original environmental scales that do not represent the true relationships between samples (those which organisms are responding to, for example – it is often the case that dose-response relationships for individuals to contaminants are more appropriate on log concentration scales). If you are still suffering agonies of indecision (!), then a purely automatic approach was given in the last section, namely to replace all variables by their ranks. This certainly achieves the twin aim of a symmetric distribution and

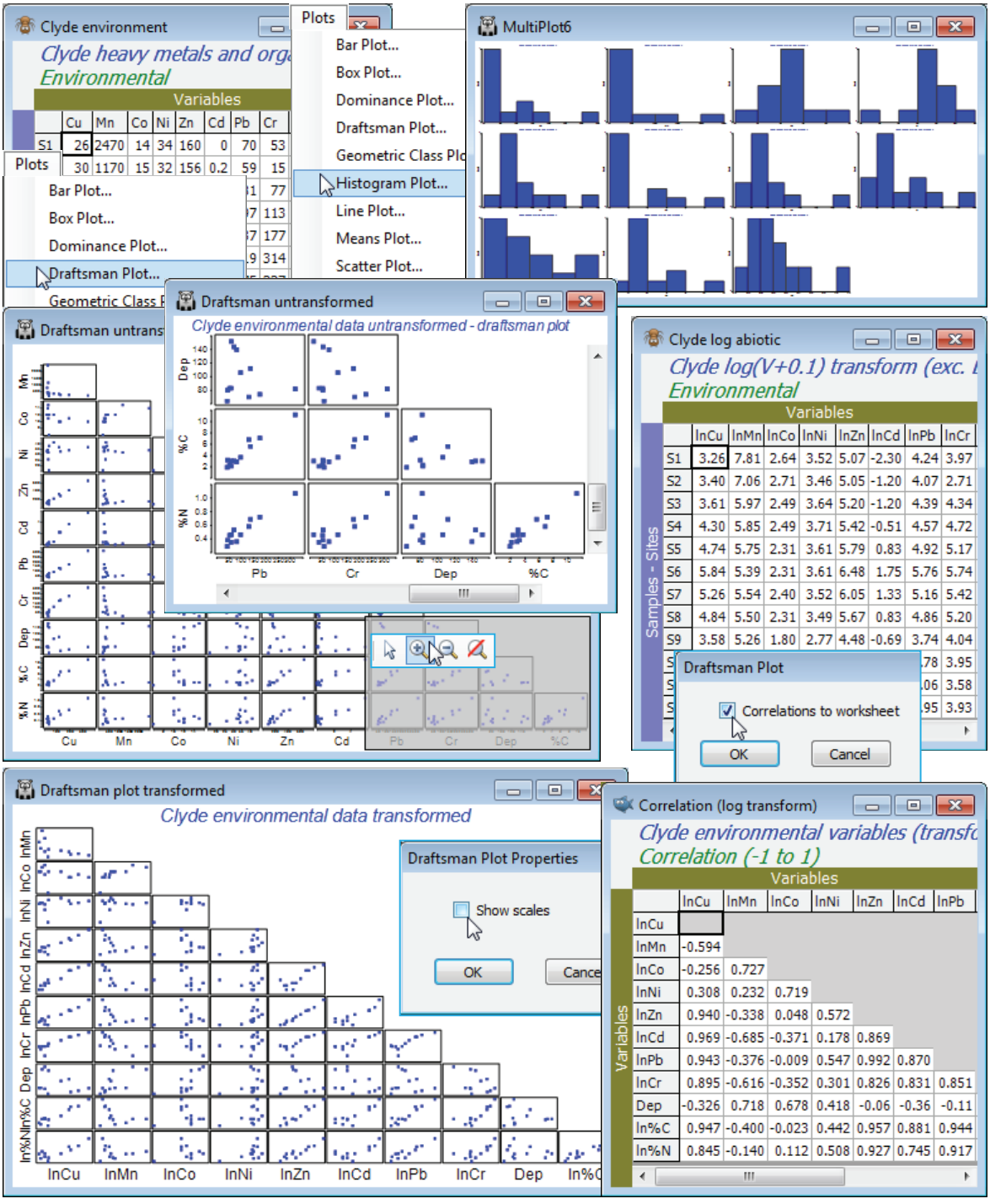
v7

v7

linear relationships (see draftsman plot below) but it must lose a little sensitivity – organisms will be responding to the dose levels themselves, on some scale, not to their rank orders!

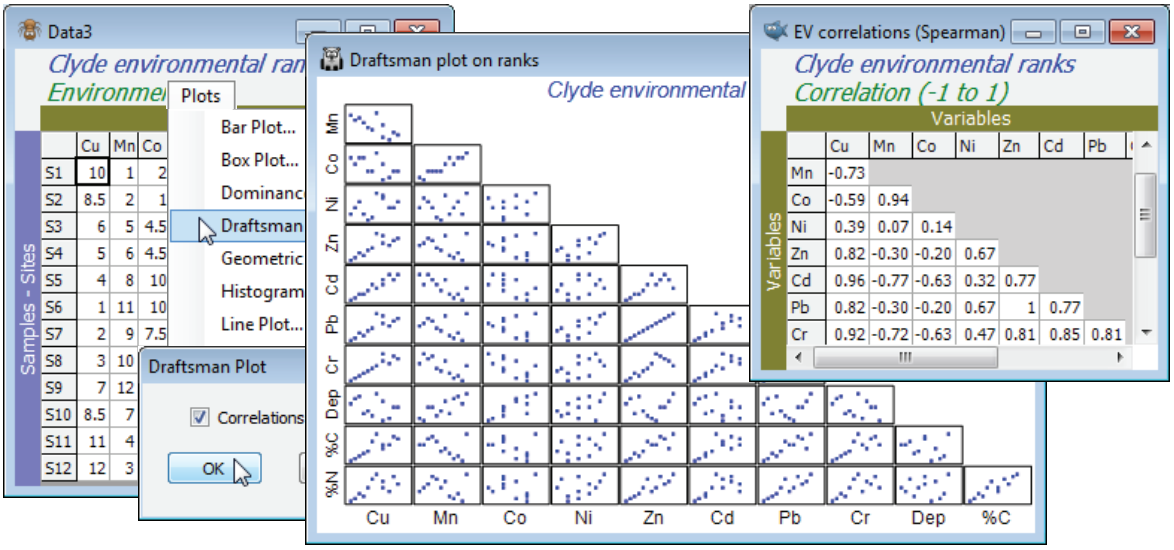
The workspace Clyde ws for the Clyde dumpground study should still be open. If not, open Clyde environment in directory C:\Examples v7\Clyde macrofauna, which has 11 environmental variables from the 12 sites (and was used extensively as an illustration in Section 11). Since there are so few samples, the draftsman plot is probably more effective here than histograms, but try both (**Plots> Draftsman Plot** and **Histogram Plot**), taking the usual graphics options to change symbol sizes, titles etc (right click then **Samp. labels & symbols** and **Titles** tab), and zooming in on part of the draftsman plot by drawing a box and **Graph>Zoom In**, or clicking the zoom icon on the tool bar. Most variables are seen to be right-skewed, which is why they were log transformed with **Pre-treatment>Transform(individual)** in the previous section (excepting water depth, a very different type of variable, which is seen to be more symmetric and not requiring transformation). Redraw the draftsman plot after you have made these transformations, this time creating the correlations among variables – more appropriate after transforming – with (✓Correlations to worksheet) in the dialog.

v7



The scales are inevitably unreadable on the full draftsman plot, so the above takes the only graphic option which is specific to draftsman plots under the **Graph>Special** menu, to turn off (✓Show scales). Keeping scales, when they are readable (e.g. under zooming), does make the point however that even in a transformed state the variables take values over different ranges, and normalising will be required (after transformation) before running a PCA. The correlation matrix shows that many of these variables are highly inter-correlated. This is not a concern for the PCA ordination which follows: part of the point of a multivariate analysis is to represent high-d data in low-d space, and this will actually be more successful if many of these variables are inter-correlated, so the points effectively lie in a 2- or 3-d subspace of the 11-d space. (It is much more of a concern for linkage methods in Section 13, which try to 'explain' assemblage structure in terms of driving variables).

The final possibility is to sidestep individual transformations altogether and work with the variable ranks (the **Tools>Rank Variables** routine covered in the previous section) – essentially this is just a different type of transformation. The variables are then forced to be symmetric, any (monotonic) relationships are certain to be linear, the variables are placed on a common measurement scale (the ranks 1 to 12 here) and there can, by definition, be no outliers – but the loss of the measurement scale is a significant drawback in using the PC axes for prediction. The correlation matrix is now of Spearman rank correlation (ρ_s) because this is ordinary Pearson correlation computed on ranks.

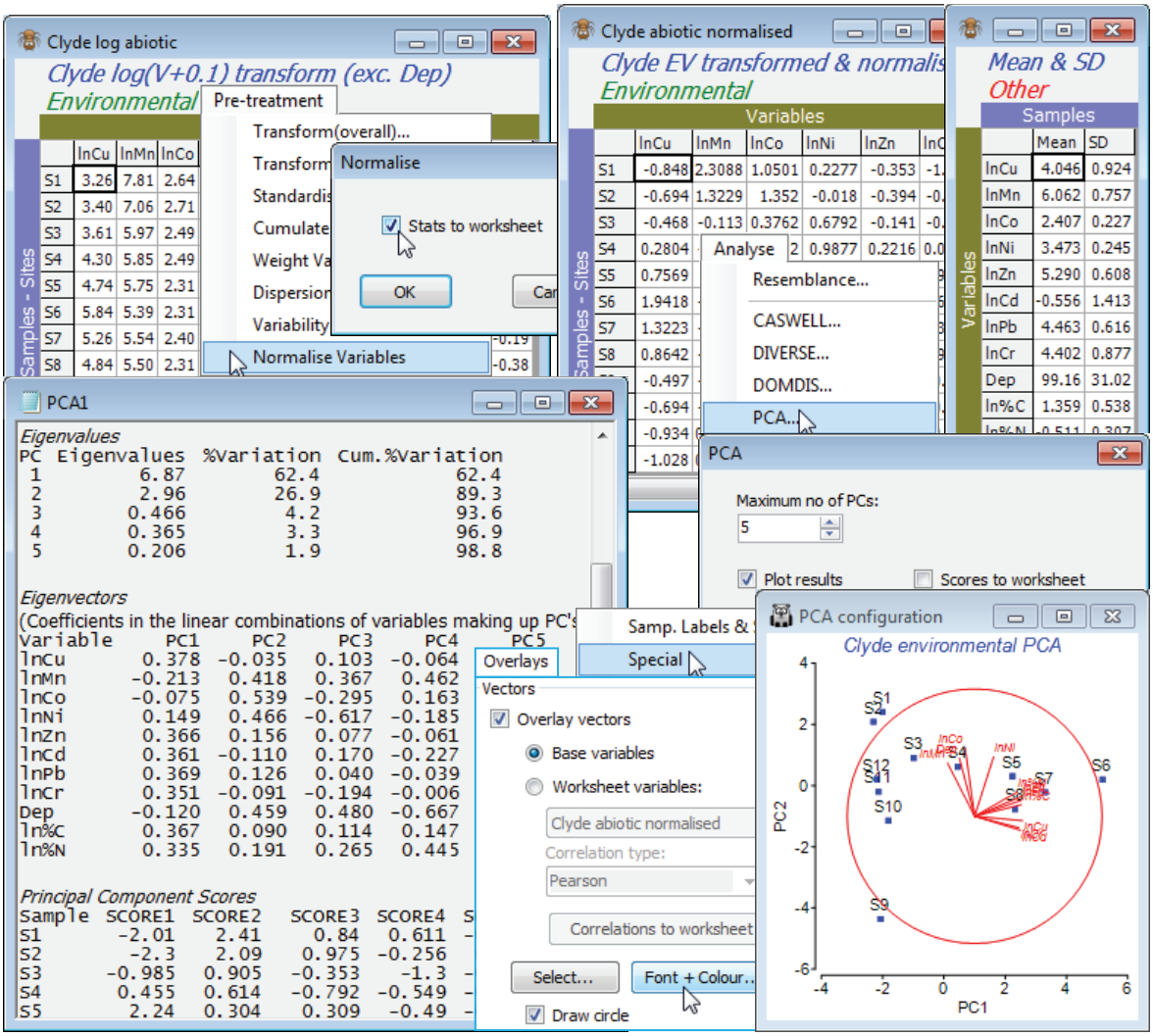


Principal Components Analysis

PCA is an ordination method in which samples, regarded as points in the high-dimensional variable space (11-d here) are projected onto a best-fitting plane, or other low-dimensional solution – the user can specify how many principal components (new axes) are required, and the routine offers 2-d and 3-d plots of any combination of these PC's. The purpose of the new axes is to capture as much of the variability in the original space as possible, and the extent to which the first few PC's allow an accurate representation of the true relationship between the samples in the original high-d space is summarised by the *% variance explained* (a percentage from *eigenvalues*). The PC's are simply a rotation of the original axes and thus a linear combination of the input variables (the coefficients are termed *eigenvectors*); PRIMER allows for superimposition of these vectors on the 2-d PCA plot. The co-ordinates of the samples on the PC axes are called the *principal component scores*, and these are output to the results, along with the %variance explained by each axis and the linear coefficients defining each PC. Chapter 4 of CiMC has a little more detail.

For the Clyde log abiotic data sheet used above, which resulted from a log(0.1+x) transform of all the environmental variables except water depth (*Dep*), take **Pre-treatment>Normalise Variables**, sending the mean and standard deviation for each of these (transformed) variables to a worksheet, and renaming the resulting data matrix Clyde abiotic normalised. On this sheet, run **Analyse>PCA**, choosing the (default) option of displaying only the first 5 PC axes, and resulting in two outputs: a detailed results window with three sections (*Eigenvalues, Eigenvectors* and *Scores*), and a PCA ordination with a superimposed vector plot (blue lines, text and circle). The vector overlay can be turned off (or changed in colour) for improved clarity, by unchecking **Graph>Special>Overlays>** (Vectors✓Overlay vectors), using **Font + Colour** to change colour from the default blue, or the circle (indicating a maximal vector) removed by unchecking the (✓Draw circle) box.

v7



PCA eigen-vector plot

Though the vector overlay has a tendency to clutter the plot, the changing contaminant load along this E-W transect of sampling sites (Fig. 1.5 in CiMC) is clear. The end points S1 and S12 lie close together and there is a strong trend from S1 to the dump centre at S6 (left to right on axis PC1), and a reversal of that trend for S6 to S12, moving away from the dump centre. The trajectory differs on the PC2 axis, however, for the two arms of the transect. The results window (heading *Eigenvalues*) shows that a 2-d PCA is a very good description of structure in the higher (11-d) space, the first axis (PC1) accounting for much of the variability (62%) and PC2 most of the remainder (a further 27%), i.e. 89% between them. The *Eigenvectors* are the linear combinations which define the axes:

PC1 = 0.378(ln *Cu*)* – 0.213(ln *Mn*)* – 0.075(ln *Co*)* + 0.149(ln *Ni*)* + …;

PC2 = –0.035(ln *Cu*)* + 0.418(ln *Mn*)* + 0.539(ln *Co*)* + 0.466(ln *Ni*)* + …,

the asterisks being a reminder that the transformed variables are normalised. It is the coefficients in these equations (eigenvectors) that the vector plot shows graphically: (ln *Cu*)* has coefficients 0.378 and –0.035, so its main contribution is to the first axis, increasing from left to right because the coefficient is large and positive, with only a slight decrease in the PC2 direction because of the small negative sign; (ln *Ni*)* has coefficients 0.149 and 0.466 so points slightly right (positive but small PC1 coefficient) and strongly upwards (large and positive on PC2), etc. The vector length reflects the importance of that variable's contribution to these particular two PC axes, in relation to all possible PC axes – if the line reaches the circle then none of that variable's other coefficients in the *Eigenvectors* table will differ from 0. The vector plot (or more clearly the eigenvector results table) show that PC1 is a roughly equally weighted combination of most of the heavy metals, Cu, Zn, Cd, Pb, Cr and organics, but not Co, Mn, Ni and Depth. The situation is reversed on the PC2 axis, with the first batch scarcely contributing at all, but the second set all increasing strongly in the positive PC2 direction. So, the first PC gives a natural way of combining the different contaminant levels into a single summary variable that characterises the main contaminant gradient.

Chapters 4 and 11 of CiMC give more on this particular example, but the principle of using a Principal Component axis as a natural, objective combination of a suite of variables is one that applies equally strongly to biomarkers, morphometric measures, water-quality metrics etc. The only difference in the latter case is that the metrics may already be standardised to a common impact scale (0 to 10, perhaps) so no prior transformation or normalisation is needed before PCA is carried out. For morphometric measurements too, transformation is often not needed and lengths, widths etc may be in common units, but normalisation may still be needed if widely different measurement ranges are involved (overall body length, setae width), to stop the larger readings completely dominating the PC's. For typical biomarker suites, transformation would need to be considered and normalisation would be essential, since entirely different scales are often involved.

PC scores

The final table in the results window is headed *Principal Component Scores* – these can instead be sent to a new worksheet by checking (✓Scores to worksheet) in the **Analyse>PCA** dialog, which facilitates their further use in PRIMER. An example would be to compute Euclidean distances among sites in PC spaces of different dimension, which could then be input to **Analyse>RELATE** (Section 14) to give a matrix correlation with the original 11-d Euclidean distances. (This is another way of measuring the fidelity of the observed low-d ordination structure to the high-d relationships, an idea we met as *cophenetic correlation* in Section 6 on the fidelity of cluster analyses, and such matrix correlations are fundamental within PRIMER, met especially in the next two sections). The PC scores are simply the x, y (or x, y, z etc) co-ordinates of the samples on the PCA plot – their values on each PC, obtained by substituting the (normalised) variable values into the above linear equations for PC1, PC2, etc. It is the ability to generate a numerical score for a fresh set of values for the same suite of variables which is one of the strengths of PCA. If values from a new site a are recorded as (Cu_a, Mn_a, Co_a, …) you can see where it fits on the contaminant scale by calculating:

$$PC1 = 0.378\{[(\ln Cu_a) - 4.046]/0.924\} - 0.213\{[(\ln Mn_a) - 6.062]/0.757\} + \ldots$$

where the means and standard deviations used in the normalisations were given in the *Mean & SD* worksheet from the normalising of the original logged data set (see output on the previous page). This is the main downside to using rank variables in a PCA, which on other grounds has much going for it – it is harder to relate new sites to the PCA from the original set of samples.

Increasing the default value of (Maximum number of PCs: 5) when running **Analyse>PCA** will print more columns of PC vectors in the results window (PC6, PC7, etc), and will allow selection of these higher PCs to be plotted in pairs or triples in the 2-d or 3-d PC configuration. However, it is rarely helpful to interpret more than the first 3 or 4 PCs, so the default computation of the first 5 is usually perfectly adequate. It is important to note that nothing changes at all in the first 5 sets of vectors if it is decided to calculate axes 6 to 10, say. Each lower-d configuration is a projection from the higher-d solution, which therefore just involves dropping out the higher axes. This is not true of MDS ordination, for which the 2-d solution is recalculated from scratch, and not just the first two dimensions of the 3-d solution.

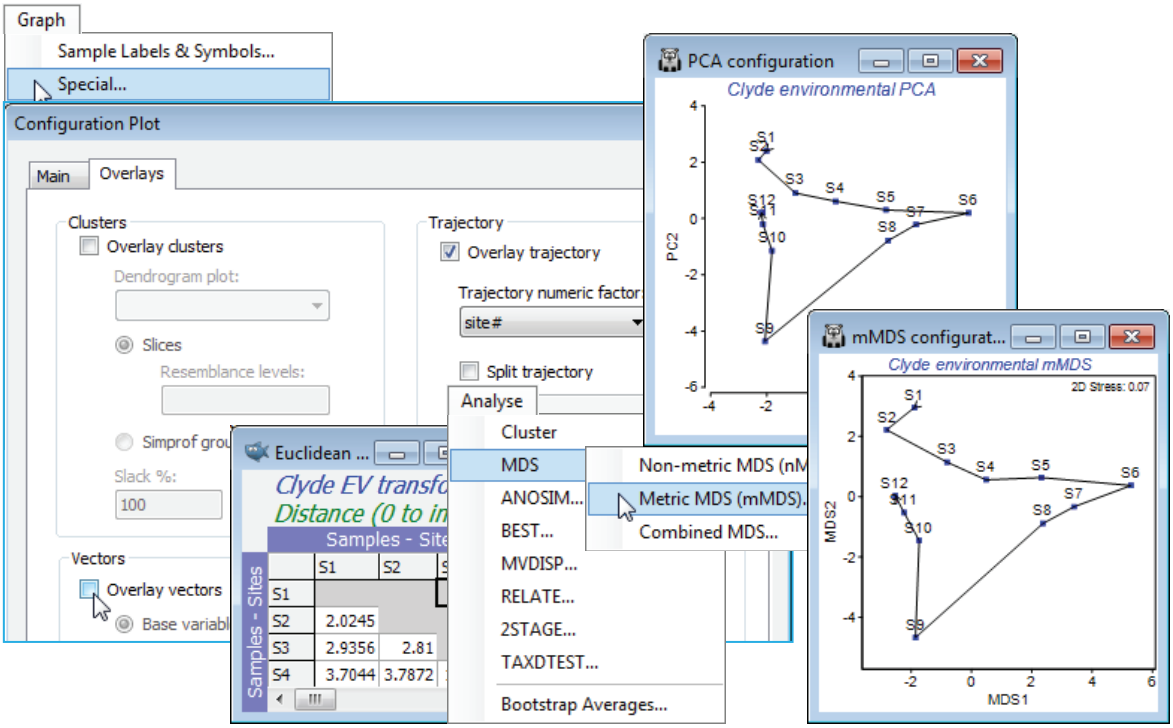
PCA plot options

Many of the options for manipulating PCA configurations are exactly the same as for MDS plots, covered extensively in Section 8, so will not be repeated – only features that differ will be shown. General rotation is not allowed in a PCA: directions have defined meanings as the axis of greatest variation, then the axis perpendicular to that with the greatest variation of that unaccounted for by the first axis, etc. However, any axis can be reflected (flipped) without affecting the interpretation in any way. Which direction the algorithm chooses to plot an axis – to the right or left, up or down, in or out etc – is arbitrary (though repeatable). In fact, in order visually to match up the PCA plot for environmental data with the *m*MDS for the same data, and the biomass (or abundance) *n*MDS ordinations, seen in the previous section, it might be necessary to run **Flip X** or **Flip Y** (or, in a 3-d plot, also **Flip Z**) either from the **Graph** or floating right-click menu. Note that when you do this, both the points and the vectors will (naturally) reverse. This does mean, however, that information already written to the results window is now slightly incorrect: the signs of the eigenvector for the axis that has been reversed need to be mentally switched (+ to – and – to +). The current location of points (PC scores) after flipping will, however, always be output correctly by **File>Save Graph Values As**, just as they are for current MDS or CLUSTER rotation states. Mention of *m*MDS raises the question as to how it differs from PCA, if both use the same metric Euclidean distances? So, now visually compare the PCA with the *m*MDS under the Ranked variables heading of Section 11.

Trajectories on PCA

From the **Graph>Special** menu, remove the vector overlay by unchecking the (✓Overlay vectors) box on the **Overlays** tab, and on the same tab, join the points along the transect with (✓Overlay trajectory>Trajectory numeric factor: Site#) – if the factor doesn't exist, create or import it, as seen under that Ranked variables heading. A better comparison would be of the current PCA with *m*MDS not on the ranked variables but on the same Euclidean distances as created from normalised and transformed variables here, so you may wish to run that **Analyse>MDS>Metric MDS (mMDS)** routine. This indicates one rather obvious difference: *m*MDS works from the resemblance matrix and PCA from the data matrix underlying that. A more important distinction is that *m*MDS does not project the points from the high-d to low-d space as in a PCA, but more carefully arranges them in order optimally to match the low-d Euclidean distance structure to the original distance matrix. Here however, all these ordination cases are effectively indistinguishable: the samples largely lie on a 2-d plane in the 11-d space making it easy for both methods to display an accurate 2-d picture.

v7



More interesting is the fact that the PCA (or *m*MDS) of the abiotic variables is an excellent match to the *n*MDS of the assemblage (also in Section 11), whether based on biomass, abundance or both, and this observation motivates the BEST routine of Section 13. (Note that a PCA of the biota is poor by comparison, since it implicitly uses Euclidean distance rather than an assemblage-based coefficient such as Bray-Curtis – and it actually fails to display a convincing species gradient even though there patently is one there. Choice of a relevant similarity is much the most crucial decision to make in multivariate analysis – a point seen again in Section 14.)

Bubble plots on PCA

Of the other options on the **Graph>Special** menu, overlaying groups from a CLUSTER run (which to be consistent must use Euclidean distance) is no different than for MDS ordination, in Section 8, and bubble plots likewise are executed in just the same way as for MDS. Though segmented bubble plots (at least for a selection of the 11 variables) will be visually clear-cut, they are not essential in this case in order to judge the contribution of individual environmental variables to a PCA derived from all of them – the vector plot provides that simultaneous information correctly for all variables. This is because the relationship of a single variable to the PC axes has to be a simple linear one, by definition of PCA, and this is the (•Base variables) option under (✓Overlay vectors), distinctive to a PCA plot. This is very different from a typical biotic *n*MDS, where the relation of single species to directions in MDS space derived from all species can often be non-linear and sometimes not even monotonic (but increasing and decreasing, impossible to represent by a vector! – Section 8). The (•Base variables) option is greyed out for MDS (or PCO) ordinations, since it only applies to PCA, but all ordinations offer vectors based on correlations with (•Worksheet variables). But the strong potential for non-linearity, e.g. of counts for a particular species across the PCA axes, makes bubble plots a much more attractive option than vectors for both PCA and other ordination types.

v7

Multiple 2-d & 3-d plots

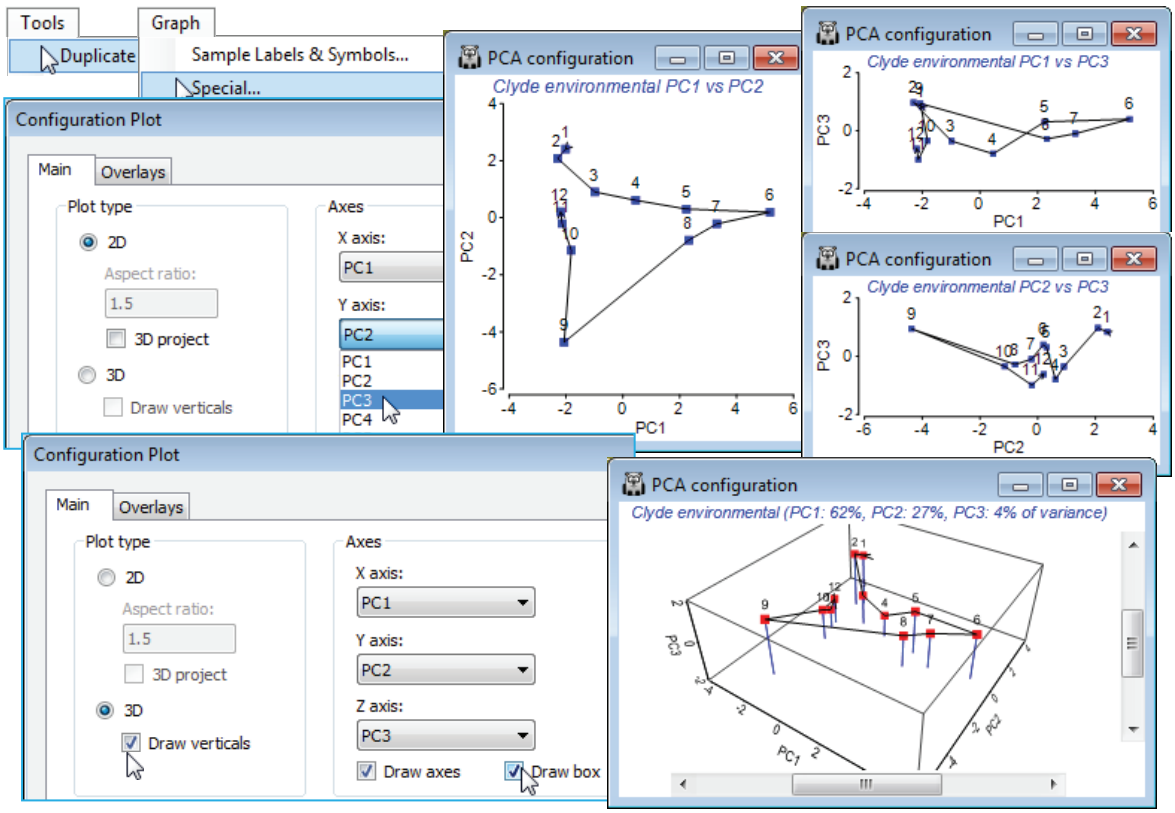
v7

As with MDS, use of **Graph>Special>Main**>Axes, with (Plot type•2D or •3D), allows any pairs or triples of axes to be plotted: (PC1, PC2), (PC1, PC3), (PC1, PC4), (PC2, PC3), (PC2, PC4), …; or (PC1, PC2, PC3), (PC1, PC2, PC4), … etc. By default, PCA is drawn with (*x, y*) or (*x, y, z*) axes rather than the full box used by *n*MDS, but either or both can be chosen – you need to select both (✓Draw axes) and (✓Draw box) to get the axis scaling and the box (the first, the second and both, are the defaults for PCA, *n*MDS and *m*MDS, respectively). Taking **Tools>Duplicate** when the active window is a plot will allow multiple copies to be displayed on the PRIMER desktop, and neatly arranged with **Window>Tile Horizontal** or **Vertical**, having first taken **Window>Close All Windows** and clicked on the series of plots to redisplay them (or the multiple plots could be placed into a new Multiplot, see Section 7). While the three 2-d plots from PC1, PC2, PC3 give, arguably, a more accurate way of publishing a static 3-d plot, the 3-d PCA graph in PRIMER is certainly the better way to view the structure on screen, and this can be manually rotated with the icon, i.e. **Graph>Rotate Axes** (rotating the data itself, within a static box – as in MDS – is not allowed since PC directions in relation to the points are fixed). Automatic rotation is with **Graph>Spin** and this can be saved as a movie file (*.mp4 or *.gif), as for MDS. **Graph>Zoom In** ( ) on a 3D plot is often a good idea, since it is usually better to see the points clearly than display all the box corners.

Interpreting PCA *vs* MDS pairwise plots

Another subtle distinction from MDS is that only a single PCA graph window is produced initially, allowing a choice between displaying a 2-d or 3-d scatter plot. This is because the PC algorithm generates just one solution, with as many PCs as requested: a 2-d PCA is just the first two axes of the 3-d PCA, etc. With MDS, the 2-d and 3-d plots are entirely separate solutions and thus held in different windows. It is possible, starting from a 3-d MDS window, to take **Graph>Special>Main** >(Plot type•2D) and generate the three pairwise plots: (MDS1,MDS2), (MDS1,MDS3), (MDS2, MDS3) – as remarked above, this gives an alternative static view of the 3-d solution, rather than an arbitrarily projected view of the 3-d box. But, unlike PCA, do not expect the (MDS1, MDS2) plot from this to be exactly the same as the purely 2-d MDS solution! They mean different things and the purely 2-d MDS solution will always be the better representation of the original relationships.

It is clear from the Clyde environmental 3-d PCA below that a 2-d ordination is perfectly adequate (noted previously from the % variance explained). The various 2-d and 3-d plots show how little absolute variation there is on the third axis – another good reason for preserving the aspect ratio, as PRIMER does for all ordinations, i.e. a distance of 0 to 2 units is the same on all axes. You may even need to change the default scaling, (✓Specify scale) on the **Z axis** tab, to (-2, 2) to get the plot below, to avoid too much compression of the PC3 axis! Save and close the Clyde ws workspace.

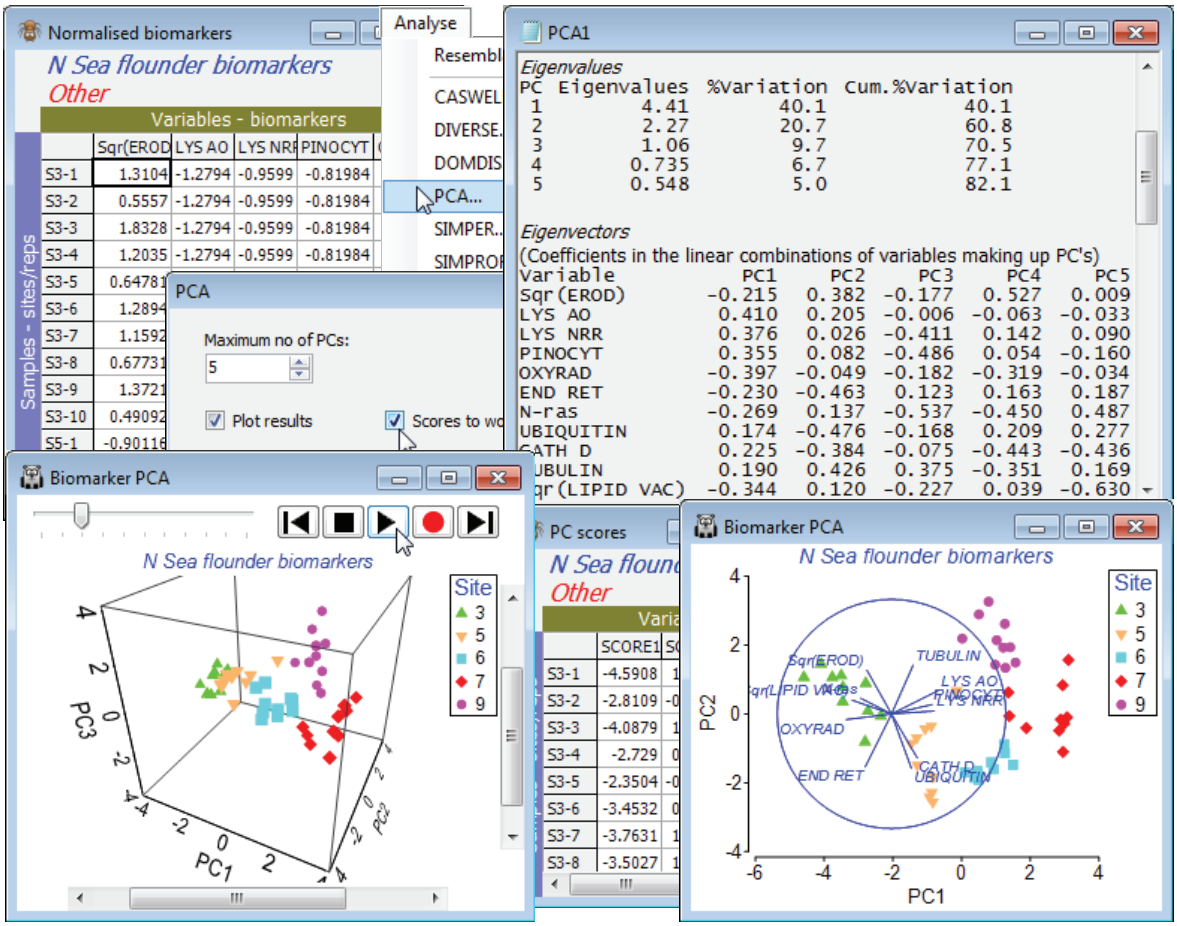


PCA of data on biomarkers

An example where a 3-d plot is marginally more necessary is given by the biomarker data last seen for a 1-way ANOSIM test in Section 9. Re-open the N Sea ws workspace, or if not available, open N Sea flounder biomarkers from C:\ Examples v7\N Sea biomarkers. Work with all the variables, not just the 6 continuous ones used in earlier sections – the remaining 5 are all ordered categorical so it is entirely legitimate to include them in a PCA (or the Euclidean distance used for ANOSIM). Previously, EROD and LIPID VAC were square-rooted with **Pre-treatment>Transformation (individual)**>(Expression: SQR(V)), but there is not much need for transforming others since there are no strong outliers (it would be pointless for N-ras which is purely binary! – though that still makes it ordered categorical). The resulting data sheet must be normalised, with **Pre-treatment> Normalise Variables**. It is rather easy to overlook the normalisation step when running PCA, but the analysis here would be disastrous without it, since the PCs are simply hijacked by the variables with highest numbers. In cases where there is a common measurement scale, normalisation may not be needed, as in the particle sizes for Danish sediments (Sections 4, 9) and Plymouth water (5).

On the normalised sheet take **Analyse>PCA,** and on the plot use the **Samp. Labels & Symbols** tab to turn off the labels, increase the symbol size and maybe change the Site key colours to avoid blue (the default colour for the vector plot). The 2-d PCA shows the separation of biomarker responses in the 5 areas, with (from the plot and the eigenvectors) sites 3 and 5 separating from 6, 7 and 9, largely on PC1, in the direction of decreasing lysosomal stability and pinocytosis, and increasing levels of oxyradicals, size of lipid vacuoles etc – indicating stress on the organisms at sites 3 and 5. What tends to separate site 7 and 9 from site 6, largely along PC2, are increased levels of EROD and Tubulin, and decreased Ubiquitin, Cathepsin D and Endoplasmic reticulum. (Remember that the vectors on the plot are read only as indicating size and direction of increase, their location being irrelevant). The eigenvectors also show that N-ras only tends to come out in the higher PC's. The eigenvalues show that 3 PCs is enough to capture over 70% of the total variability (a good target figure), so it is worth a look at the 3-d plot with **Graph>Special>**(Plot type•3D) & (Axes✓Draw box). Turn off vectors from the **Overlays** tab, by unchecking (✓Overlay vectors) and **Zoom In**, **Rotate Axes** and **Spin** from the right-click or **Graph** menu. The 3-d plot certainly separates the sites clearly but the extra 10% of explained variation in comparison with the 2-d plot does not alter the interpretation to any extent. Resave the workspace as N Sea ws and close it.

v7



13. Linking assemblage to environment (*BEST*: Bio-Env, *LINKTREE*)

BEST rationale

The main rationale for the **Analyse>BEST** procedure in PRIMER is to find the best match between the multivariate among-sample patterns of an assemblage and that from environmental variables associated with those samples. The extent to which these two patterns match reflects the degree to which the chosen environmental data 'explains' the biotic pattern. This leads naturally to the idea of searching over subsets of the abiotic variables for a combination which optimises that match, namely the *best* explanatory variables – see Chapter 11 of CiMC for details of the method. The concept is a more general one (see also Chapter 14), and BEST can equally be used to find: subsets of taxa which best match a fixed environmental data set (e.g. vulnerable and opportunist species characterising a known impact gradient); subsets of biota which best match a different biotic matrix (e.g. key coral species which may be structuring a reef fish community) or even the same biotic matrix (e.g. a small subset of species, perhaps chosen from a set of easily-identified taxa, which generates the same multivariate sample pattern as would the full assemblage). Parallel applications for different data types can also be envisaged, for example: a subset of tissue contaminant levels that best 'explain' a suite of biomarkers, or conversely, a subset of biomarkers that best identify a body burden contaminant gradient; a subset of geomorphological variables that best characterises an existing classification of rivers or coasts; a small set of morphometric or genetic/molecular measures that is as effective as a larger set in discriminating two putative species, and so on.

Bio-Env *vs* BVStep

BEST amalgamated the earlier (PRIMER 5) BIOENV and BVSTEP procedures (hence BEST = Bio-Env + Stepwise) since they had an identical purpose – to search for high matrix correlations, rank-based, between a fixed sample similarity matrix (typically from a species assemblage) and resemblance matrices generated from different variable subsets of a supplied data matrix (usually a transformed and normalised suite of environmental variables presumed to include those 'driving' the assemblage structure). The only difference in operation is that BIOENV carries out a complete search of all possible combinations of variables from the datasheet, whereas BVSTEP caters for the common situation in which there are too many variables to do an exhaustive search, and a forward-stepping and backward-eliminating stepwise procedure is necessary to arrive at a (possibly) optimal set. Within **Analyse>BEST**, the first choice is therefore of Method•BIOENV or Method•BVSTEP. (BVSTEP will be discussed in Section 14, where it becomes essential for use on biotic matrices).

Change to active sheet for BEST

v7

In what is one of the very few examples of 'moving the furniture around' between PRIMER 7 and earlier versions, the active window for a run of **Analyse>BEST** is no longer the data matrix of (usually abiotic) variables, from which selections are made to best match to a fixed resemblance matrix (usually of assemblage pattern), but the converse, i.e. **Analyse>BEST** is run from an active sheet which is the fixed resemblance matrix, and a secondary matrix supplied which is the (abiotic) data matrix from which variables are selected. There are several good reasons for this switch, the most compelling of which is that it allows much greater consistency in the way PRIMER decides which samples to use in an analysis when there is only a partial match of sample labels between the two data sets. The consistent rule now, throughout PRIMER (and PERMANOVA+), is that the active matrix (in its currently selected form, if any selections are in place) determines that sample set. Any secondary matrix supplied to the routine (here, the abiotic variables) are treated as a 'look-up' table from which the required set of samples is extracted. Thus, the environmental matrix can (and sometimes will) cover a much wider range of sites than are utilised in the current community samples – they might for example be interpolations from some physico-chemical or remote-sensing model for the whole region. What is required is that BEST can find all (biotic) resemblance sample labels in the (abiotic) data matrix, otherwise an error is returned – with the usual relaxation of strict label matching that if the two matrices have exactly the same number of samples, i.e. BEST will ask the user if it should should proceed on the assumption that the samples are in the same order. [Other benefits from switching the active matrix for BEST include consistency with the DISTLM routine in PERMANOVA+, which is the semi-parametric equivalent to the non-parametric BEST program, and a close multivariate analogue of (univariate) multiple linear regression. In standard statistical thinking, the *response* here is the community sample, thought of as subject to sampling/ spatio-temporal variability and that is regressed on the observed values of the *explanatory* variables (the latter considered fixed, under a conditionality argument). PERMANOVA+ routines thus start from the response, as given by the community resemblances – which are always the active sheet – and explanatory variables (in DISTLM), covariates (in PERMANOVA) etc are always secondary.]

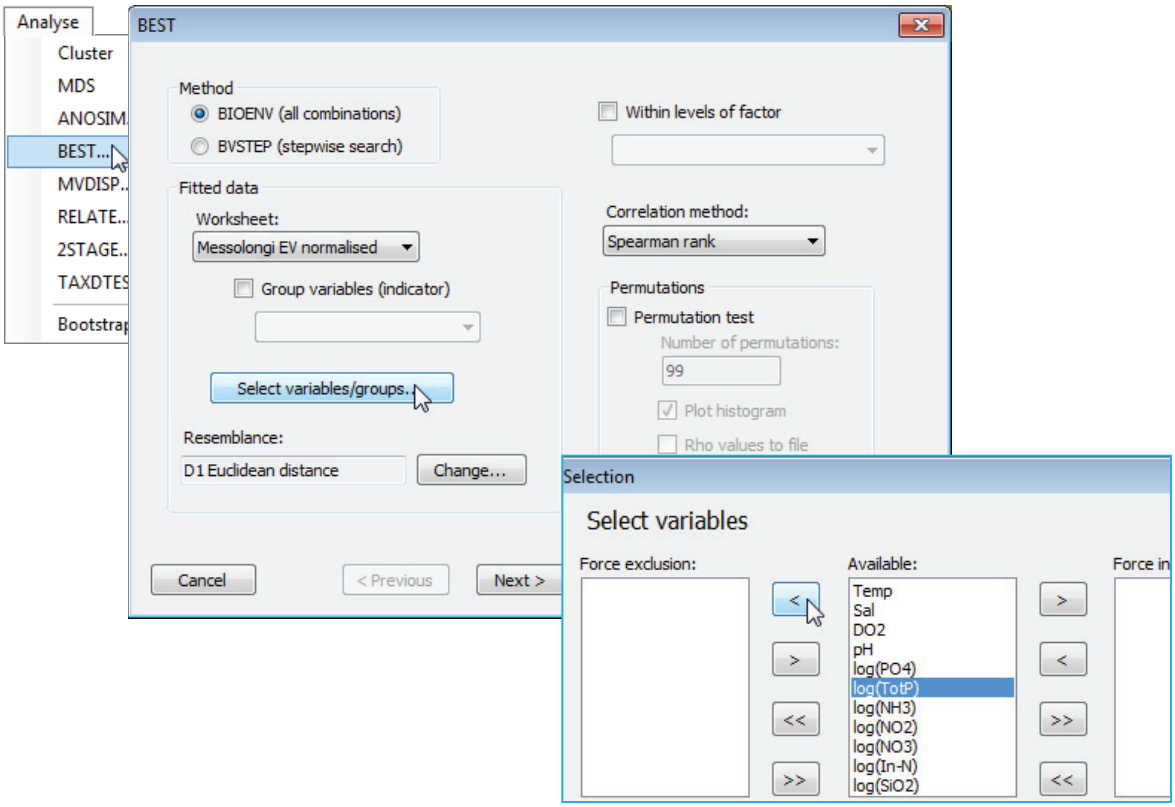
Grouping variables in BEST

v7

After the initial choice of Method, the next area on the BEST dialog inputs the explanatory (*fitted*) data worksheet and, in a new option in PRIMER 7, allows the user to specify an indicator for that sheet which groups its variables into indivisible sets. For example, one 'explanation' of differences in community structure might simply be geographical location. If that is a latitudinal gradient it could be represented by a single explanatory variable, but if the samples have 2-d rectangular co-ordinates (or even 3-d), then it would not make sense for the BEST search procedure to consider explanations which included the *x* but not the *y* co-ordinate. If this pair of variables are specified as a group – they have the same level of the indicator – then they will be selected (or not) as a single unit. (Whether it is sensible to include geographic location in the explanatory set for a community pattern is another matter altogether, since location cannot be causal *per se* – an organism does not know its GPS co-ordinates! Instead, the model is more likely to be that of species responding to other enviromental variables which are changing with geographical position, and those should be the variables in the explanatory set). Other examples of grouping might be to separate variables of different types: Valesini FJ *et al* 2014, *Estuar Coasts* 37: 525-547 give a BEST analysis of this sort in which estuarine fish communities are related to groups of variables representing wave exposure, substrate type, marine water intrusion etc, in order to determine if one or more sets of variables are particularly influential – Chapter 11 of CiMC gives slightly more detail.

Selecting variables & resemblance

After the (✓Group variables(indicator)) check box, the next option is a **Select variables/groups** button, which gives the usual type of selection dialog with three panes. The default is for all the variables – for which read 'groups of variables' if the previous check box is ticked – present in the (Fitted data worksheet:) to be displayed in the (Available:) pane. These will then be picked and dropped in all combinations. Variables that are moved to the (Force exclusion:) pane will never enter any of the combinations considered, e.g. you might choose to exclude a variable which is very highly correlated with another in the list. Those variables in the (Force inclusion:) pane will be included in every combination, e.g. you might know that a particular environmental variable is causal for the assemblage, and therefore always want to include it when considering whether adding other variables improves the 'explanation'. The choice of (Resemblance:) coefficient for the explanatory variables then follows. The default for this is determined by the datasheet type – often environmental, and thus Euclidean distance – but can be altered to any of the numerous measures which PRIMER offers, through the **Change** button. Importantly, for environmental variables on different scales, the supplied explanatory variables worksheet should be in its normalised form before **Analyse>BEST** is run – there is no option within the dialog box to add this pre-treatment step before selection of Euclidean (or other) distance measure.



2-way BEST

v7

On the right of this main dialog box for **BEST** is another option new to PRIMER 7, also covered in Chapter 11 of CiMC, namely the check box (✓Within levels of factor). Essentially, this gives a *constrained* (or *2-way*) BEST procedure in which the match in sample patterns between (usually) abiotic variables and the assemblages is calculated separately for each level of the supplied factor, and the appropriate matching statistic (a matrix correlation) averaged over those levels. Selection of the variables is made simultaneously in all levels of that factor, and the optimum match is therefore given by the variable set which succeeds in maximising this averaged matrix correlation. The idea is that there may often be situations in which the dominant differences between communities are due to an (unordered) categorical factor, which cannot be simply accommodated by adding another (ordered) variable to the abiotic matrix, and is perhaps a nuisance factor in trying to understand the detailed relationship between abiotic and biotic patterns – its effect is fully removed by matching only within the strata of this factor. The analogy with 2-way ANOSIM is strong, e.g. removing the effect of Site when testing for differences over Time, by constructing an R statistic for a Time test separately for each Site, and averaging them. So this analogous matching procedure can be thought of as a *2-way* form of BEST. Just as with ANOSIM, it may be possible (and sensible) to run BEST entirely separately within each stratum of the nuisance factor, e.g. match abiotic to biotic patterns completely independently for each of a small number of geographical regions. However, where there are rather few samples in each region, 2-way BEST provides a 'half-way house' in which matching is carried out separately for each region but with common choice of the abiotic variable set, which makes sense if there is not a strong *interaction* between the effect of an environmental variable and the region (e.g. an interaction would be when salinity is crucial to the community in region A but, though varying equally greatly in region B, has no effect on the community structure there). Under these (*additive*, not interactive) conditions, such a constrained, 2-way BEST routine may lead to a much more incisive (powerful) analysis.

The BEST matching statistic, ρ

v7

On the mid-right of the main dialog for BEST, the box headed (Correlation method:) now offers three non-parametric choices and one parametric correlation: Spearman rank, Weighted Spearman rank, Kendall tau and Pearson, covered in equations (11.3), (11.4) and (2.3) respectively in CiMC. These are the measures of agreement (matching statistics) between the two resemblance matrices, e.g. biotic and abiotic, and correlations (ρ) are calculated by matching element to element. The logic is that if the true driving abiotic variables are selected, and two sites have very similar suites of values for these, then the assemblages will also be very similar (and vice-versa), so the triangular matrix elements should rank in the same order. Ranks are usually appropriate not only because of their central role in PRIMER, underlying a non-metric MDS ordination and the hypothesis testing procedures in ANOSIM and RELATE, but also because the two resemblance matrices may use entirely different coefficient types, e.g. Euclidean distance in (0, ∞) and Bray-Curtis in (0,100). Whilst the above logic then leads one to expect a monotonic relation between their values, there is no reason to expect a linear relationship between a distance and a finite-range dissimilarity, so standard Pearson correlation will generally be less effective. However it is included in PRIMER 7 to cover situations in which, for example, two sets of Euclidean distances, or two sets of Bray-Curtis dissimilarities are being matched, and a standard correlation may then be more acceptable. Though weighted Spearman was constructed to be more relevant to this specialised case of matrix correlations (rather than standard rank correlations of two variables with independent entries), in practice there is rather little to choose between the three rank-based coefficients.

v7

Limiting the number of combinations

The final area of the main BEST dialog, headed Permutations, which carries out the *global BEST test* for statistical significance of the best matching combination of variables, is deferred until later in this section. The Next > button, under (Method•BIOENV), gives a dialog with a single entry, a choice of (Max num of trial variables/groups:), for which 5 is the default. This limits the search to ≤5 (abiotic) variables at a time, and this maximum number should be increased, where feasible. A default of the total number of input variables is not used because the number of combinations of these in an exhaustive search could be very large: for p variables there are $c = 2^p - 1$ combinations, and a practically realistic limit therefore has to be about $p = 17$ (giving $c \approx 100{,}000$). The context for Section 13 is the matching of subsets of environmental variables to assemblage patterns. Quite often, the number of abiotic variables is then <17 or, if not, the number should probably be pruned before running BEST – so only a full search (BIOENV) will be illustrated now. BVSTEP could be run in much the same way on a larger set, but the reason this is likely to prove unattractive is that,

with so many abiotic variables, it is inevitable that they will be strongly inter-correlated. There are then a plethora of equally good solutions and a rather unfocussed interpretation. Deletion of all but one of a highly mutually-correlated set of variables and/or prior reduction to one representative of each different type of environmental variable, may be desirable, just as in multiple linear regression (see the discussion in Chapter 11 of CiMC). In some of the other applications – e.g. when the data matrix is of species variables and *a priori* selection defeats the point of the analysis – the stepwise form (Method•BVSTEP) will be essential, and such an example is seen in Section 14.

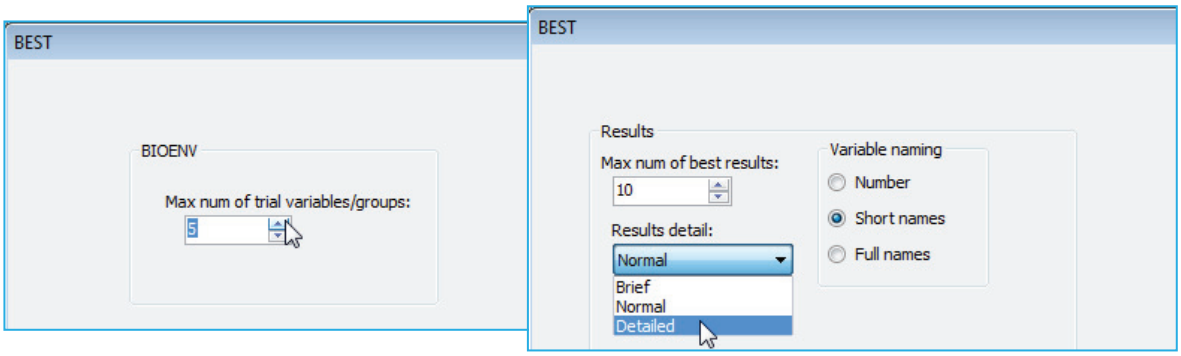
BEST results detail

v7

The Next > button now takes you to a Results dialog box, which controls two features of the BEST results window, the quantity of output and how the explanatory variables are identified. For a run of BIOENV, (Results detail: Brief or Normal) will output just two short tables, headed *Best result for each number of variables* and the overall *Best results*. The former is actually the more useful (and this particular summary of the full output is new to PRIMER 7). It identifies the best single variable, which on its own has the highest (matrix) correlation ρ with the biotic resemblances, then the best pair of explanatory variables, the best triple, and so on. The second table simply selects the variable combinations which overall give the highest correlations seen, in decreasing ρ order (and irrespective of the number of variables in those combination), the total number of them listed being determined by the (Max num of best results:) box, with default 10. (Results detail: Detailed) is probably preferable initially – until you become accustomed to the BEST procedure – because it outputs not just the above summary tables but prefaces them by the ordered decreasing values of ρ for all variable combinations, organised as in the first summary table by groups of the same number of variables. This again has the three columns: number of variables, matching coefficient ρ, and the variables used. (The distinction between Brief and Normal is used only in BVSTEP, where there is an extra level of results from the different random starts of the search procedure – see Section 14).

v7

Previous versions of PRIMER used only the variable number in these tables, with a list at the start of the results relating those numbers to the variable labels. This made the large tables cumbersome to interpret, so PRIMER 7 offers three options: variable naming by (•Number), (•Short names) or (•Full names). The last of these is the full variable label and the middle option a truncated form, each variable with as few of its initial characters as are sufficient to make the names distinct. This much improves the readability of the output, but there are occasions when it is still desirable to re-run BEST with numbers, so that the best set can be easily selected from the original matrix with **Select>Variables>**(•Variable numbers:), copying and pasting the number string to this box.



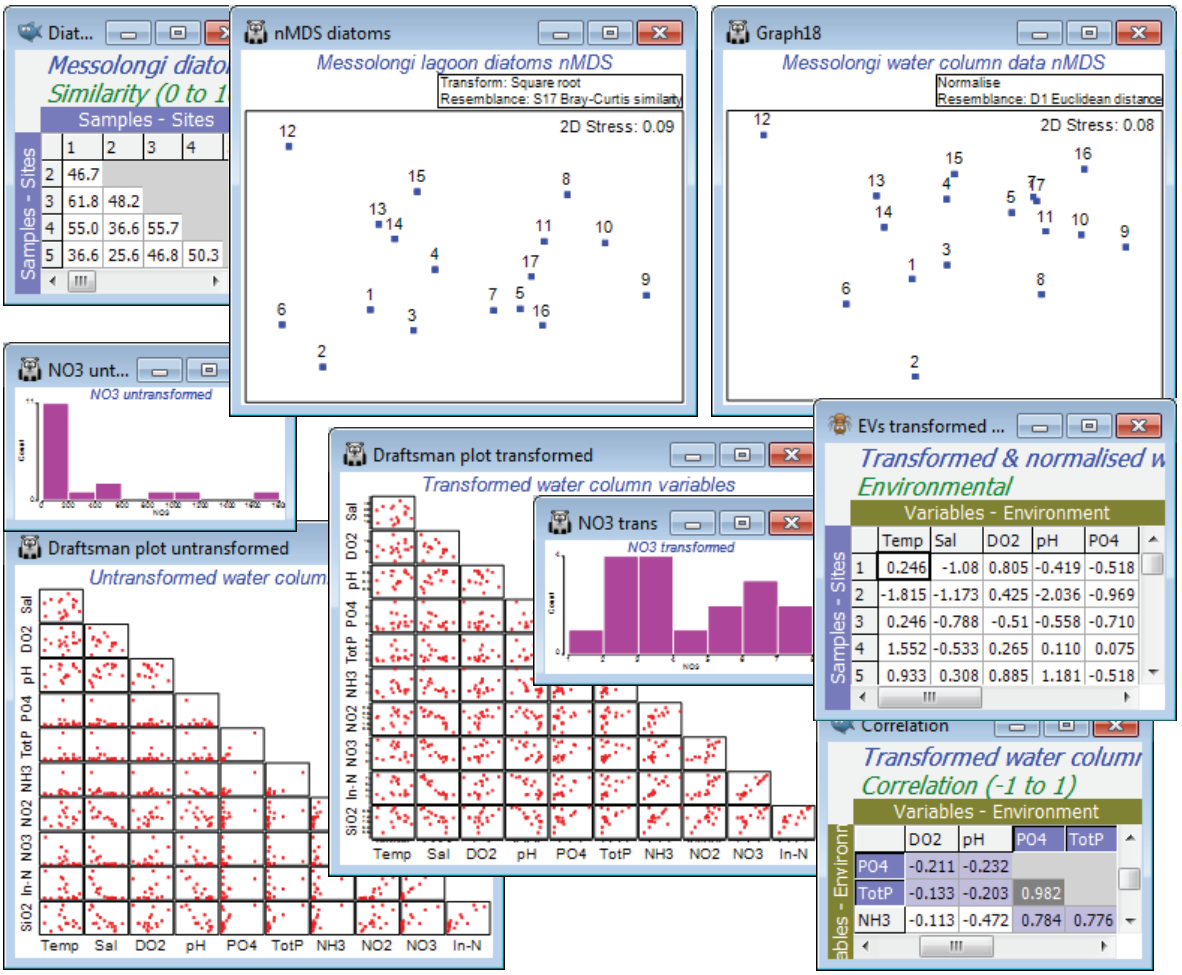
(Messolongi diatoms & abiotic data)

A study of diatom assemblages (abundances of 193 species) at 17 sites in the lagoons of Messolongi, Aitoliko and Kleissova in Eastern Central Greece was undertaken by Danielidis DB (1991), *Ph.D. thesis, Univ Athens*. At each site, a suite of 11 water-column data was also recorded: Temperature, Salinity, DO_2, pH, PO_4, Total P, NH_3, NO_2, NO_3, Inorganic N and SiO_2. The data files are Messolongi diatom density and Messolongi environment in C:\Examples v7\Messolongi diatoms. This is an ecological study of how the diatom communities relate to the water-column variables.

v7

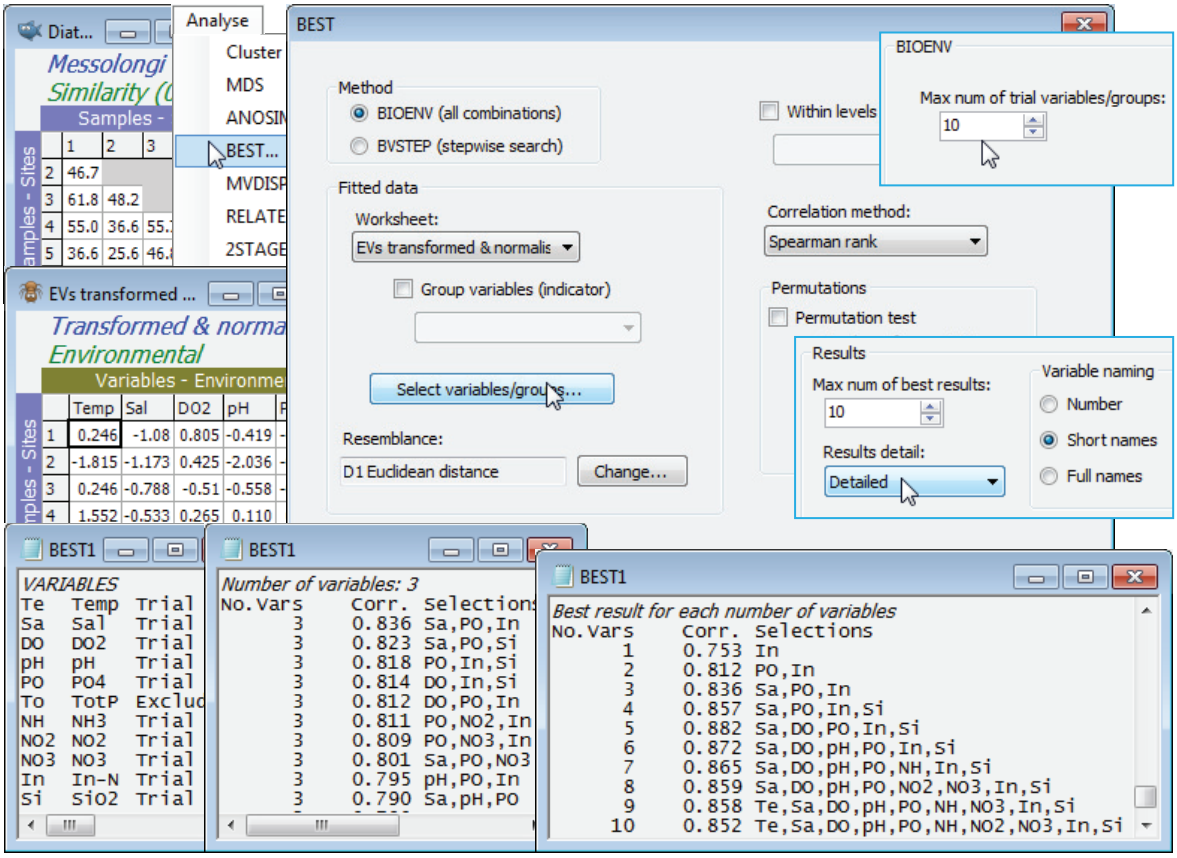
Square-root transform the abundance file and take Bray-Curtis resemblances, plotting the *n*MDS as usual. **Plots>Draftsman Plot** or **Histogram Plot** show that a log transform would be desirable on the nutrient concentration variables *PO_4, TotP, NH_3, NO_2, NO_3, In-N* and *SiO_2*, but *Temp, Sal, DO_2* or *pH* do not need any transformation. As in the previous section, carry this out by highlighting (not selecting) the variables to be transformed and take **Pre-treatment>Transform(individual)>** (Expression: log(V)), unchecking the (✓Rename variables) box – readability of the BEST output is

improved if not all the variable names look like *log(...)*!, so bear in mind that PO_4 means $log(PO_4)$ etc, from now on. Re-running **Draftsman** and **Histogram Plots**, and also taking (✓Correlations to worksheet) for the former, shows that the distributions now have greatly reduced right-skewness. Two variables, PO_4 and *TotP* are seen to be strongly collinear, and it will make sense to drop one of them in the **BEST** run – they are, in effect, the same variable. You can pick out which are the very strongly correlated variables by **Select>Samples>**(•Values>0.95) on the correlation matrix produced by the draftsman plot – and potentially repeat again with (•Values<-0.95), though there are none of the latter here. This will display only those rows and columns of the triangular matrix with a value >0.95 somewhere, just PO_4 and *TotP* in this case. On the transformed data, take **Pre-treatment>Normalise variables**, and the among-sample relationships, in terms of these 10 abiotic variables, can then be seen either by **Analyse>PCA** directly on this matrix or calculating Euclidean distance and putting that into MDS. As expected, since both are based on Euclidean distance, the two ordination methods for the abiotic data give very similar 2-d plots but more remarkable is the near-perfect match of biotic and abiotic analyses – the 193-species diatom community is highly predictable from knowledge of these 10 water-column variables.



In fact, the match is even better with fewer abiotic variables. With the diatom resemblance matrix as the active sheet, run **Analyse>BEST>**(Method•BIOENV) & (Worksheet: EVs transformed & normalised), forcing exclusion of TotP under the Select variables/groups button, with the default of Euclidean resemblance and (Correlation method•Spearman rank), and leaving the Permutation box unchecked. On the Next > dialog, increase to (Max num of trial variables/groups: 10), since all 1023 combinations will run in a reasonable time. On the final dialog, (Results detail: Detailed) and (Variable naming•Short names). The results window and particularly the summary table of *Best results for each number of variables* shows that ρ is maximised (at 0.88), for the 5 variables: *Sal*, DO_2, PO_4, *In-N*, SiO_2 and slowly decreases beyond that, as more variables are added. The best 3-variable solution (*Sal*, PO_4, *In-N*) does nearly as well (ρ = 0.84), and on the principle of parsimony might be preferred as a simple 'explanatory' set of abiotic variables for these diatom communities. Causality, of course, is not established – see the comments in Chapters 11 and 12 in CiMC.

v7



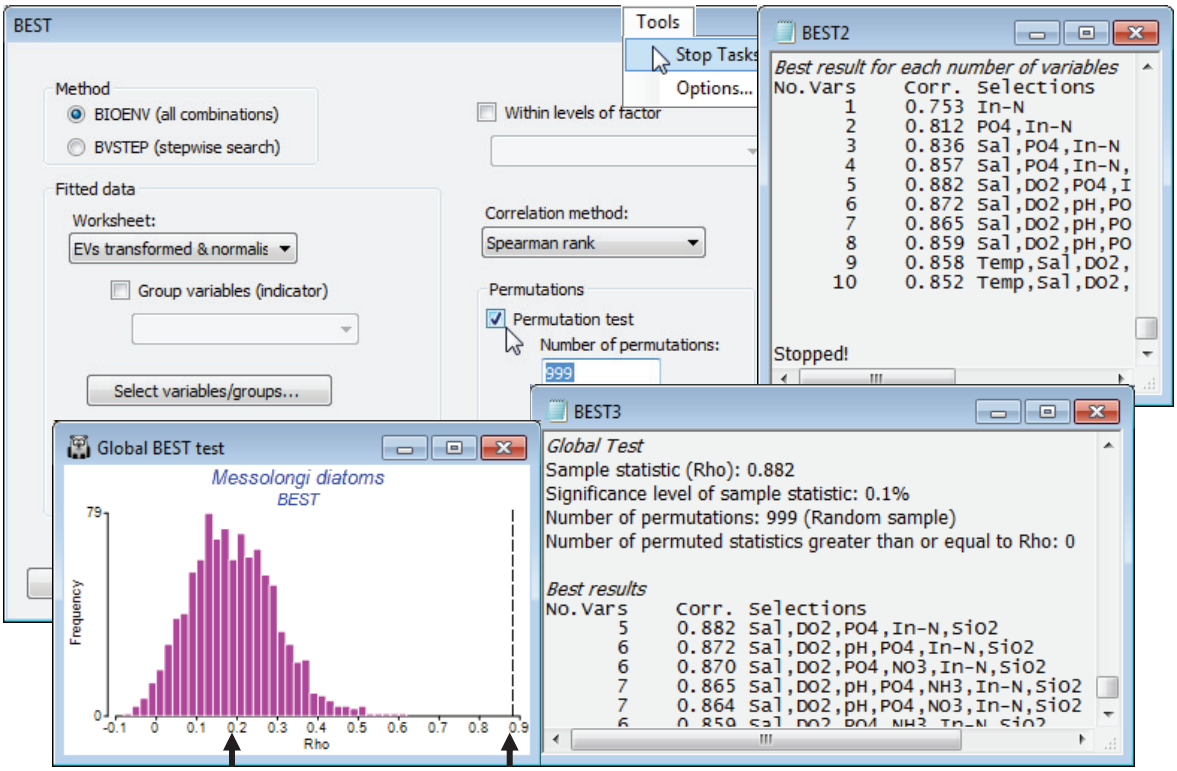
Global BEST test

The question of statistical significance testing on the results of the Bio-Env (or BVStep) procedure naturally arises. Section 14 describes **Analyse>RELATE**, a (non-parametric) form of Mantel test. For any two independently-derived resemblance matrices, defining the relationships among the same set of sample labels, one can use permutations to test the null hypothesis H_0: *no agreement in multivariate pattern*. The measure of agreement is the (usually rank) correlation coefficient ρ, discussed above, between the corresponding elements of the two triangular arrays, with ρ = 0 representing the null hypothesis. The ρ values that it is possible to observe by chance, if the null hypothesis is really true, can be generated by randomly permuting one set of sample labels relative to the other (thus destroying any real link) and recalculating ρ, over many random permutations. RELATE could therefore be applied to testing agreement between an assemblage and the full set of environmental variables for the same sites (though not for all other linkage problems mentioned earlier, e.g. between the full assemblage and a subset of conspicuous species, since independence is violated – any subset of species will bear *some* relation to the full set). It is important to realise, however, that RELATE cannot be applied to the subset of environmental variables that result from a run of BIOENV: these have been selected precisely to maximise the matching coefficient ρ with the assemblages. Even where there is no real match, the optimum ρ produced by BIOENV will inevitably be >0. We need a test which allows for this selection bias, and this is the *global BEST test* (Clarke KR *et al* 2008, *J Exp Mar Biol Ecol* 366: 56-69, and Chapter 11, CiMC), a permutation procedure accessed on the first dialog box from **Analyse>BEST**. The idea is simple: randomly permute one set of sample labels in relation to the other, then run through the full BIOENV (or BVSTEP) process to generate the best match ρ. Another permutation of the labels is then generated and the BIOENV run repeated again, and so on (for 99 times by default, because of the intensive computation involved – but preferably more). This produces 99 values of ρ in a histogram, which represents the null hypothesis. The real ρ is compared with these, as for any PRIMER permutation test – if it is larger than any of them, then the null hypothesis can be rejected at p<1% significance. Actually, this is the sole example in PRIMER of a statistic (ρ) which does not take the value 0 for the null hypothesis – as indicated above, the mean ρ is certain to be >0 under H_0.

v7

When the new 2-way BEST routine is run, by taking the option (described earlier) to remove the effect of a categorical variable by matching only within its levels, the test proceeds in the same way but with a constrained permutation of biotic sample labels within – not across – those levels.

From an active sheet of the lagoon diatom resemblance matrix, re-run **Analyse>BEST>**(Method• BIOENV) & (Worksheet: EVs transformed & normalised) with most options as before but this time taking (Permutations✓Permutation test)>(Number of permutations: 999) & (✓Plot histogram), and (Variable naming•Full names). On a slow machine, or with more samples than here, you will probably need to reduce the number of permutations to 499, or 199, or 99. The latter is adequate if the result is clear cut, but results in a much less smoothed histogram, and you will wish to calculate more in borderline cases. Remember that you can use always use **Tools>Stop Tasks** (or the icon on the Tool Bar) to interrupt a permutation test that, from observing the green progress bar, is *v7* clearly going to take too long – note that since it computes and outputs the BEST results tables for the real data before embarking on the random permutations, you will not lose these if you stop the routine prematurely. An alternative is to multi-task, carrying on with other PRIMER activities as the permutation test runs in the background – this is not a problem. In addition to generating a null *v7* distribution histogram (for which you can change the bin size, colours etc with **Graph>Special** as usual), the test adds a small section to the results window, headed *Global Test*, whose format is as for the ANOSIM test, Section 9. It gives the real value of ρ and its % significance, 100×(1+(no. of permuted ρ≥observed ρ))/(1+no. of perms). The real ρ (0.88) is well to the right of the upper tail of the null distribution, p<0.1% (i.e. P<0.001). Note also that the mean of the histogram is not zero but around ρ = 0.2. The strong selection pressure, over a large number of variable combinations, is able to produce an artefactual match up to about ρ = 0.4 or even 0.5, though there is no question that the null hypothesis is rejected here – such a good match of water column indices to the diatom assemblages, as seen in the earlier biotic and abiotic MDS plots, clearly cannot be due to chance. A final step would be to select only the BEST set of abiotic variables and repeat the Euclidean MDS.



Linkage trees – rationale

Another technique for linking sample patterns based on assemblage data to a suite of environmental (or other) explanatory variables was also discussed in Clarke KR *et al* 2008 *J Exp Mar Biol Ecol* 366: 56-69 (see the last topic in Chapter 11, CiMC). The well-established statistical procedure of *Classification And Regression Trees* (CART) was further developed in an ecological context by De'ath G 2002, *Ecology* 83: 1105-1117, termed *Multivariate Regression Trees* (MRT). PRIMER implements a modification of this, in a form which is consistent with the non-metric philosophy underlying the rest of the package. The connection with regression is minimal (and confusing) so the more descriptive term *linkage trees* is used by PRIMER for its variation of the procedure. Its *v7* real affinity is with Cluster analysis (Section 6, under heading Binary divisive clustering), and it is therefore accessed in PRIMER v7 by running **Analyse>Cluster>LINKTREE**. In fact, it is a form of *constrained* binary divisive clustering in which the successive divisions of the full set of biotic

v7 samples, seen in the *unconstrained* divisive clustering of **Analyse>Cluster>UNCTREE** (Section 6), are limited to those splits of each group (into two new sub-groups) which have an explanation in terms of larger or smaller values of a specific explanatory (typically abiotic) variable – consistently so on either side of that divide. In other words, all constraints are a threshold inequality on a single abiotic variable and this set of inequalities form the possible 'explanation' for the biotic structure.

We have already seen two techniques for linking assemblage patterns to abiotic variables: bubble plots (Section 8) and the above BEST procedure. BEST has the advantage of looking at the abiotic variables in combination, trying to identify a subset which is sufficient to 'explain' all the biotic structure capable of explanation, and the matching procedure takes place in the full high-d space, i.e. on the respective resemblance matrices. But on its own, this falls short of a full interpretation because it does not demonstrate which variables take high or low values for which samples. Bubble plots give the latter but are only satisfactory where the low-d biotic *n*MDS has acceptable stress as an approximation to the full biotic pattern. *Linkage trees* can fill this gap: they can take the subset of abiotic variables identified by BEST, and use them to describe how the assemblage samples are optimally split into groups (in the high-d space), and interpret this, e.g. Group 1 communities have Salinity<23ppt but Group 2 are from >26ppt (with no samples between these salinity thresholds). Group 1 and 2 samples are then each divided into two by a different threshold on the same abiotic variable, or more likely by a different abiotic variable. The result is divisive clustering of the biotic samples, and an environmental interpretation, e.g. for the lagoon diatoms, the cluster of sites 13,14, 15 below has (Salinity<23), (54<PO_4<82) and (In-N<965), the only sites to meet those conditions.

Non-metric, non-linear, non-additive

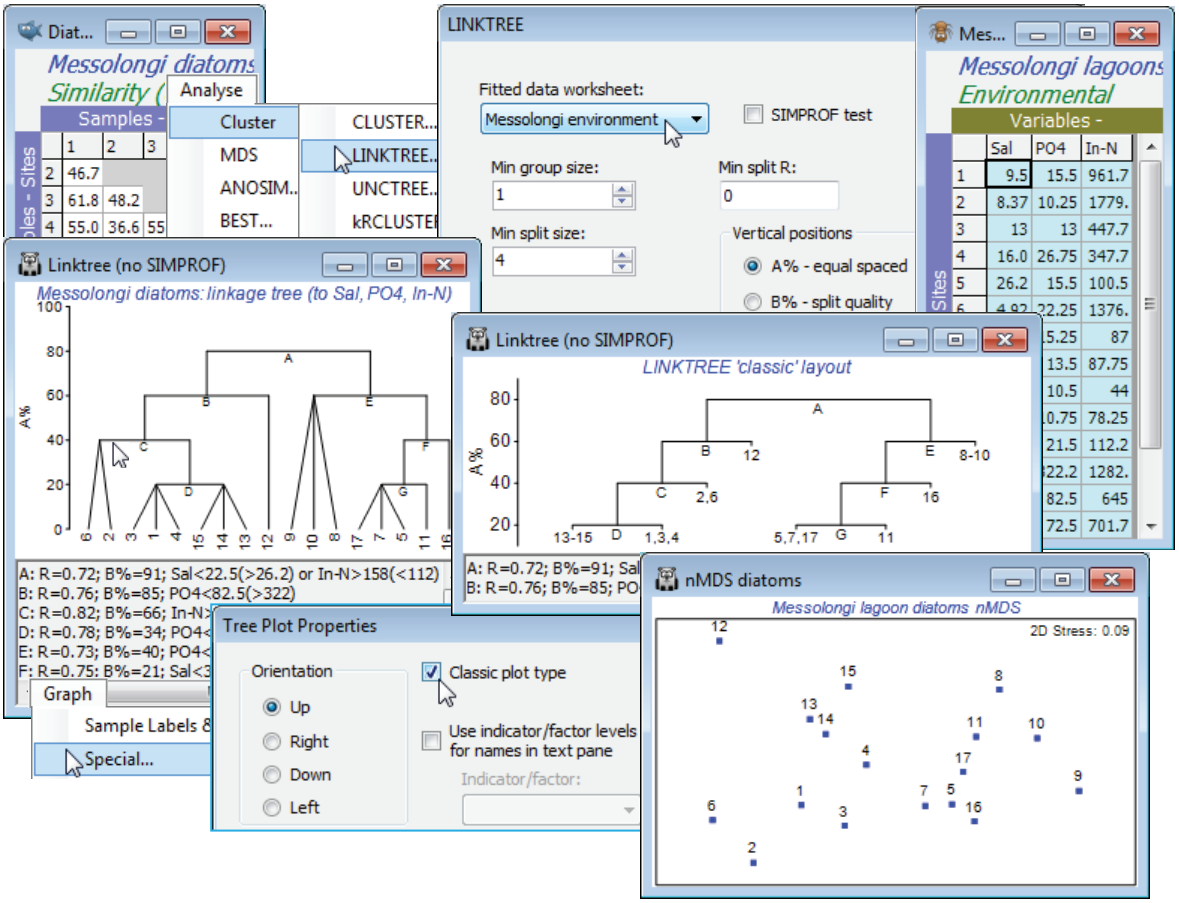
The **Analyse>Cluster>LINKTREE** routine has a number of features that are designed to mesh to the PRIMER approach. Firstly, as seen for unconstrained UNCTREE clustering (Section 6), each successive split of the biotic samples into two groups (of potentially unequal size) maximises the ANOSIM R statistic (Chapter 6, CiMC). An ANOSIM test is not carried out, of course (that would be totally invalid since the same data would be used to define the groups as to test them!) but R has a general role as a non-parametric measure of multivariate difference between groups (in high-d),

v7 rather than just as a test statistic. Unlike the much more computationally intensive UNCTREE, not all possible binary divisions are permitted (there are ~2^{16} possibilities for just the initial split of the 17 lagoon sites, which is why UNCTREE needs an iterative search algorithm!). In fact LINKTREE can simply examine all splits that correspond to a threshold condition on an abiotic variable (so for 3 variables there are at most 3×16=48 ways to divide 17 samples into two groups). Secondly, the procedure is truly *non-metric*, not just on the community resemblance matrix but also on the abiotic variables. A (monotonic) transform of the environmental variables can make no difference to the outcome of LINKTREE, since all that is being used is how a criterion like In-N<965 or >1380 splits up the samples (again there are no samples with In-N between 965 and 1380). That division is unchanged under transformation, just becoming log(In-N)<log(965) or >log(1390) for example. Thirdly, and more subtly, the way the different abiotic variables are combined in the partitioning of the biotic samples is clearly *non-linear* but is also *non-additive*. In contrast, BEST is non-metric and can certainly accommodate non-linear responses of the assemblages to driving environmental variables, but does make an implicit assumption that their effects are additive. For example, if high PO_4 were to be an important variable in separating the diatom communities but only in low salinity environments (with equally large variation in PO_4 having no effect on the biota in high salinities), then this would clearly degrade the BEST match (ρ). Such interactions are one explanation for the failure to get a good match, along with several others: high sampling 'noise', failure to measure the important abiotic variables, communities structured by competition not external driving variables, etc. However, LINKTREE attempts only local explanations – rather than holistic ones in the way BEST does – and is clearly capable of showing, for example, that PO_4 is important for structuring low-salinity groups but not high-salinity ones (with similar PO_4 ranges). A big disadvantage with the local (piecemeal) explanations offered by LINKTREE is that many abiotic inequalities will explain the same assemblage divisions, unless the environmental variable set is initially drastically pruned. An advantage is that it is geared towards prediction, and not just interpretation.

LINKTREE (Messolongi lagoons data)

v7 Continuing the lagoon diatom study, having first selected (highlighting then **Select>Highlighted**) the optimal 3-variable set (*Sal*, *PO_4*, *In-N*), from the above BEST run, in Messolongi environment, and again with the diatom resemblance matrix as the active sheet (not the abiotic data, as in earlier PRIMER versions), take **Analyse>Cluster>LINKTREE>**(Fitted data: Messolongi environment)

v7 & (Min group size: 1) & (Min split size: 4) & (Min split R: 0) & (Vertical positions•A%), and uncheck the (✓SIMPROF test) box for this run. These conditions determine that a group of size 3 will not be divided but that groups of size 4 or more will be, if R exceeds 0 (though a minimum split value of 0 effectively means that this last condition will never come into play). Such stopping rules are arbitrary and inferior to SIMPROF tests, seen next. (Note that since transforms change nothing, the original form of the abiotic matrix is preferred, for ease of interpreting the scales. Of course, normalisation is not required either, since abiotic variables are no longer combined).



The output is a tree diagram with a text pane below it. The first split (A) in the assemblage data is between sites 1-4, 6, 12-15 (left hand side of the biotic MDS plot shown above, from earlier) and 5, 7, 8-11, 16, 17 (right hand side) – a very natural divide in the ordination (though remember that the procedure works in the high-dimensional space not the 2-d MDS). This has ANOSIM R = 0.72. It is characterised by low or high salinity (Salinity<22.5 to the left, and >26.2 to the right). Note that the inequalities in the text pane (repeated in the results window) are always in this order, with the branch to the left first and the branch to the right following, in brackets. It follows that if the tree is rotated, by clicking on a horizontal bar exactly like a CLUSTER dendrogram, then the inequality in the text pane reverses (the reason for a dynamic text pane rather than just a static results window). Alternatively, the same split A of samples is obtained by choosing In-N>158 to the left and <112 to the right. R is the same whichever variable is used, of course – both can 'explain' that biotic split so both are reported. Moving down the left of the tree, the next split (B) divides sites 1-4, 6, 13-15 from site 12, with an R of 0.76, on the basis of PO_4 (high phosphate at site 12). Then C splits 2, 6 from 1, 3, 4, 13-15 at R = 0.82, again with two explanations (and convincingly on the MDS), etc. The end result is 8 groups of sites, each determined by a series of abiotic inequalities. Note that R has no tendency to decline/increase on moving down the tree – the ranks are recalculated for each new subset of samples. An absolute measure (B%) which does generally decline with finer group
v7 distinctions was given in Section 6 for the analogous UNCTREE plot. This can be used as a *y* axis for the plot (see over); the A% scale just displays arbitrary equal-spaced steps but that can help the clarity of the 'classic' form of linkage/regression tree plot, which is also shown above, and is an option on the **Special** menu for linkage plots. This menu also allows the plot to be re-oriented, as for a dendrogram, and can replace long variable names in the text pane with short indicator levels.

SIMPROF test in LINKTREE

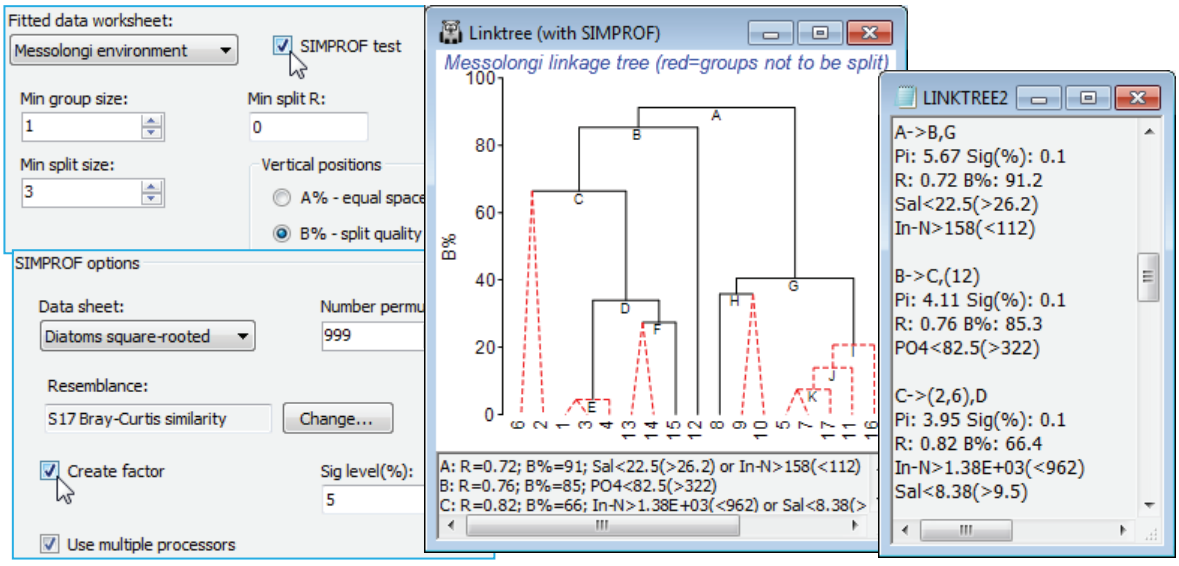
Low values of B% correspond to samples which are rather close together on the MDS plot and the question naturally arises as to whether these samples should be split at all – is there any evidence that the biological assemblages differ among the sites 5, 7, 11, 16, 17? If not, then we should not be seeking an environmental variable which distinguishes two subgroups within them. The SIMPROF test (Section 6) answers this question and provides a statistical basis for interpretation of a further subdivision. The test is the same as used with the unconstrained cluster analyses of Section 6 – the real profile of the biotic resemblances, in rank order, is compared with many repeated profiles from randomly permuting species values across these 5 samples, separately for each species. The test statistic π measures departure of the real profile from the mean of the random profiles, and this is set against the range of values it takes for the deviation of (further) random profiles from this mean. A large real π implies significance, e.g. if it is larger than all but 49 of the 999 random profiles then homogeneity of the assemblages in this group would be rejected at p≤5%, and it is justifiable to interpret the next division LINKTREE makes – the text pane and results window continue to list all divisions permitted by the other stopping rules but the tree branches in red are not significant and it would be unwise to interpret those splits. The results window gives SIMPROF π and p values and a factor is created of the SIMPROF groups which can be used to show those groups on an MDS, say.

v7

Run **Analyse>Cluster>LINKTREE** as before on the diatom resemblances, this time taking (Min split size: 3) so this criterion does not enter – remember SIMPROF can never split a group of two – and (Vertical positions•B%) & (✓SIMPROF test). Look at the entries on the SIMPROF options dialog, but you will probably not need to change any. Since the test is on the biotic data not the environmental, the program steps back in the Explorer tree to find the default (Data sheet: Diatoms square-rooted) whose rows are to be permuted, and the (Resemblance:) specified will be the one used for the active matrix (Bray-Curtis here). You may need to reduce the number of permutations for much larger data problems (this intensive routine exploits available multi-core processing) or just run LINKTREE without SIMPROF tests, and do some selective tests on a few key splits with **Analyse>SIMPROF** on these selections in Diatoms square-rooted. The plot here shows that (5,7, 11,16,17) do not differ (π ≈ 0.95, p<35%) but (1,3,4,13-15) do differ (π ≈ 2.3, p<1%) and are split into three interpretable groups. Note also the uneven steps (large and small group differences) in the B% scale, which is now comparable across branches, unlike the equi-spaced A% scale.

v7

v7



Missing data in linkage trees

Note that LINKTREE is able to tolerate some missing data in the abiotic matrix – the piecemeal form of LINKTREE's conclusions lends itself to analysing whatever complete matrix is available locally, i.e. within each created subdivision. But distortions in interpretation from unavailability of explanatory variables in some sets of samples and not others are almost inevitable. A final point to make is that it is always interesting to compare a constrained **Cluster>LINKTREE** with the unconstrained, but otherwise very similar, **Cluster>UNCTREE** tree structure. Here, exactly the same divisions are found (and of course confirmed in the same way by the SIMPROF tests). Where there are major differences, this suggests that natural clusters in the samples are not being well identified by the current abiotic suite, perhaps because a key variable is missing (though there are many other possible reasons! – see the discussion on *reversals* in B% plots in Chapter 11 of CiMC).

v7

14. Further matching of multivariate patterns (*RELATE, 2STAGE, BEST + MVDISP*)

RELATE on resemblance matrices

The BEST routine in the previous section introduced the concept of measuring how closely related two sets of multivariate data are, for a matching set of samples, by calculating a rank correlation coefficient (Spearman's ρ, Kendall etc) between all the elements of their respective (dis)similarity matrices. Thus, if the among-sample relationships agree, in exactly the same way in both data sets (e.g. the two closest samples are 3 and 5, the next two closest are 7 and 15, …, and the furthest apart are 6 and 11), then the rank correlation ρ = 1, a perfect match. (These element-by-element correlations of two resemblance matrices are known as *matrix correlations* or *Mantel coefficients*, *v7* though Mantel – working in epidemiology – defined them with standard Pearson correlations, a less flexible option than rank correlations for our purposes but one which PRIMER now provides). The two resemblance matrices to be compared in this way need not be of biotic and environmental data respectively, but can come from any source: biotic compared with biotic, abiotic with abiotic, biotic with a *model matrix*, etc – it is only necessary that they refer to matching sample labels.

PRIMER performs the calculations by the **Analyse>RELATE** routine, with active window as one of the resemblance matrices to be compared. In fact, RELATE allows the user either to supply the second matrix as another triangular resemblance sheet (the general case) or to specify one of two special cases of simple model matrices, which the routine then constructs for itself. The first is referred to as *seriation*, where the data is compared to a linear sequence, either in space or time, i.e. the matching coefficient ρ assesses the extent to which samples follow a simple trend: adjacent samples being the closest in species composition, samples two steps apart the next closest, and so on, with assemblages from the first and last samples differing the most. Chapter 15 of CiMC gives more detail on model matrix construction, and draws the clear link between the RELATE test for *v7* seriation and the *ordered ANOSIM* test seen in Section 9 (and described in Chapter 6 of CiMC). RELATE, however, is able to accommodate more complex hypothesised models than the simple serial trends of ordered ANOSIM (with or without replication), e.g. the other model RELATE constructs automatically is simple *cyclicity*, with the sample relationships thought of as matching those of distances between points placed equidistantly around a circle. A possible context could be monthly samples taken over a full year. With a seasonal signal one might expect adjacent months to be the most similar, months two steps apart less similar etc, but the assemblage structure for later months gradually returns to that at the start, so that Dec and Jan are only one step apart, not 11.

Model matrix construction

Model matrices corresponding to more complicated structures than simple seriation or cyclicity need first to be constructed by the user and then entered to RELATE in the same way as any other resemblance matrix being matched to the active sheet. There are at least three ways of obtaining such model matrices. Firstly, they can be read in directly as a triangular matrix, e.g. as an existing physical distance matrix between the sampling points – there the idea would be to judge how well the community dissimilarities match geographical layout. Secondly, they can be produced from simple *x* (or *x, y* or *x, y, z*) co-ordinates of the sample points by running this 1- (or 2- or 3-) variable data sheet through **Analyse>Resemblance**, choosing Euclidean distance. For example, if simple seriation (perhaps for an inter-annual time trend) was not already catered for directly in **Analyse> RELATE**, it could be handled by creating a data sheet with one variable and *n* samples, of entries 1, 2, …, *n*, and calculating Euclidean distances – producing a lower triangular matrix with 1's on the diagonal, 2's on the first off-diagonal, …, down to *n*–1 in the lower left corner. And a model distance matrix corresponding to a monthly season cycle would result from the *x, y* co-ordinates of numbers on a clock face being input to Euclidean distance (again, non-normalised). This will not give model entries which are integers but the distances will be in the correct rank order – which is all that matters for RELATE's rank correlations). For a geographical layout, enter the metric form of lat/long co-ordinates to Euclidean distance. Thirdly, however, PRIMER helps you to construct model matrices directly from specified factors using **Tools>Model Matrix**, which is run when the active sheet is the biotic resemblance matrix to be compared with the model. An example given below is of *seriation with replication*, namely four groups of samples considered to be at points 1, 2, 3, 4 along a line (thus dissimilarity between group 1 and 2 is less than that for 1 and 3, or 2 and 4, and that for 1 and 4 is larger still). This cannot be handled by choice of the *seriation* option in **RELATE** because that is only appropriate to single samples at each space (or time) point – here there are replicates in each group, considered to be at distance 0 from each other. **Tools>Model Matrix**, specifying a numeric factor with appropriate levels 1, 2, 3, 4, will create the correct model.

RELATE hypothesis test

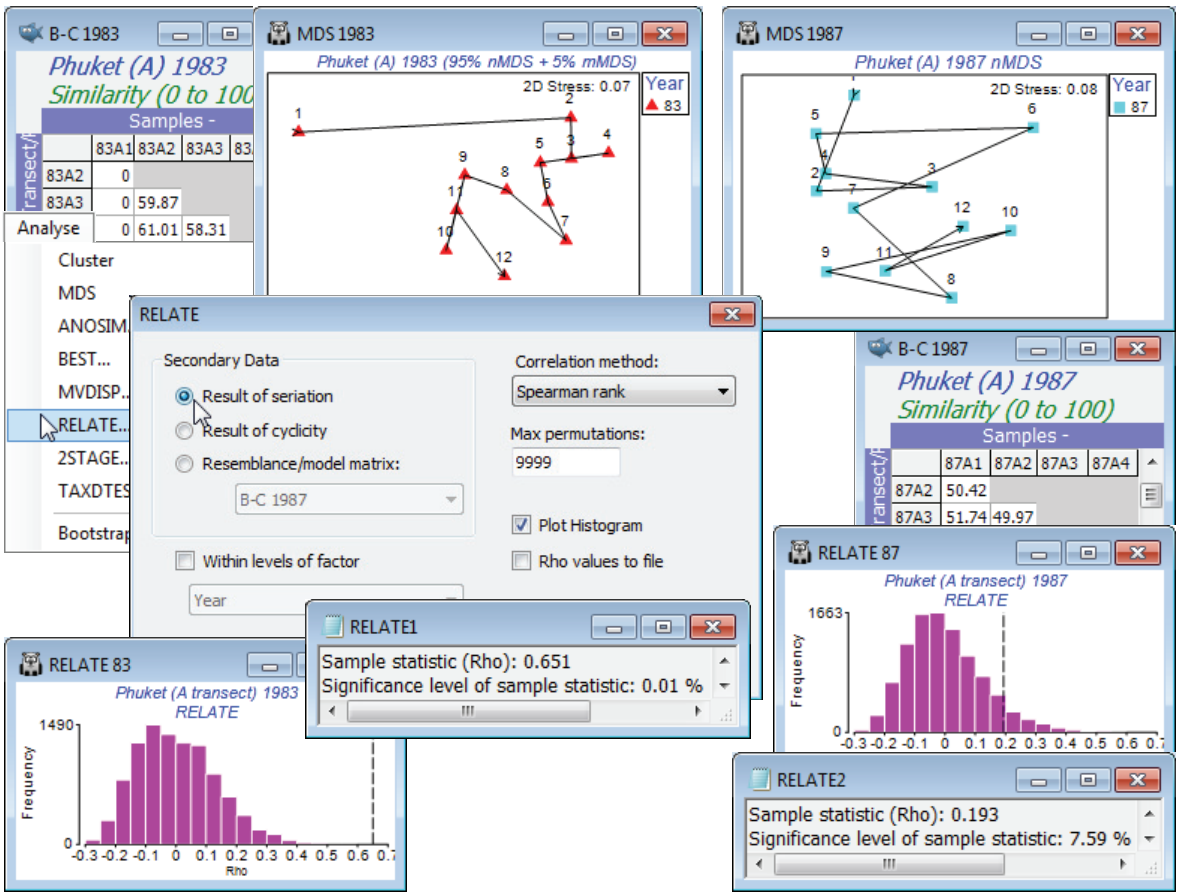
A permutation test can be applied to the matching coefficient ρ between any two resemblance matrices which are independently derived, with all sample labels in the active matrix matched with (some) labels in the supplied resemblances. As remarked in the previous section (in the context of testing for a significant match between biotic composition and a suite of environmental variables), it would not be appropriate to use RELATE on two matrices derived from the same data, e.g. by different transformation or aggregation level on the same set of species abundances. Under the null hypothesis of no relation in sample structure between the two similarity matrices, $\rho \approx 0$. The null distribution of ρ either side of zero can be obtained by randomly permuting, many times, one (or both) sets of sample labels and recalculating ρ, to derive a histogram with which the true value of ρ can be compared. The following example is given in Chapter 15, CiMC (Breakdown of seriation).

Seriation (Phuket coral transects)

The Phuket coral-reef assemblages at equi-spaced positions down an onshore-offshore gradient (transect A) from Phuket Island, were seen previously in Sections 8, 9 and 11. Open the workspace Phuket ws, or if not available open just Phuket coral cover 83-87 from C:\Examples v7\Phuket corals, square-root transform, calculate similarity and create *n*MDS plots such as that of Section 8, separately for the two years 1983 and 1987. Do this by selecting the 12 samples along the transect in 1983, with **Select>Samples>**(•Factor levels)>(Factor name: Year>**Levels**>(Include: 83) – note that **Select** works in just the same way on a resemblance matrix as a data sheet – then take **Tools> Duplicate** to make a copy of this smaller resemblance matrix, renaming it B-C 1983. **Analyse**> **MDS>Non-metric MDS** with default options, except (✓Fix collapse)>(Metric proportion: 0.05). This is needed to avoid the collapse of the *n*MDS plot because of the outlying first point on the transect (as seen in Section 8). With **Graph>Special>Overlay>**(✓Overlay trajectory)>(Numeric trajectory factor: Position), the serial change in coral community over the transect positions is clear. Repeat these steps for 1987, giving resemblance B-C 1987. The choice of (✓Fix collapse) is not necessary here but if you run the *n*MDS with and without this option you will see that it makes no difference at all to the outcome – the metric proportion of the minimised combined stress function is so small that it cannot influence the plot unless there really is no non-metric information to use, as for Position 1 in 1983, when the metric stress kicks in. (You may want to use the Procrustes routine, **Graph>Align Graph** on one of these plots, specifying the other as the (Configuration Plot:) to match to – see Section 8 Align graphs automatically – to see they are indeed identical).

v7

v7



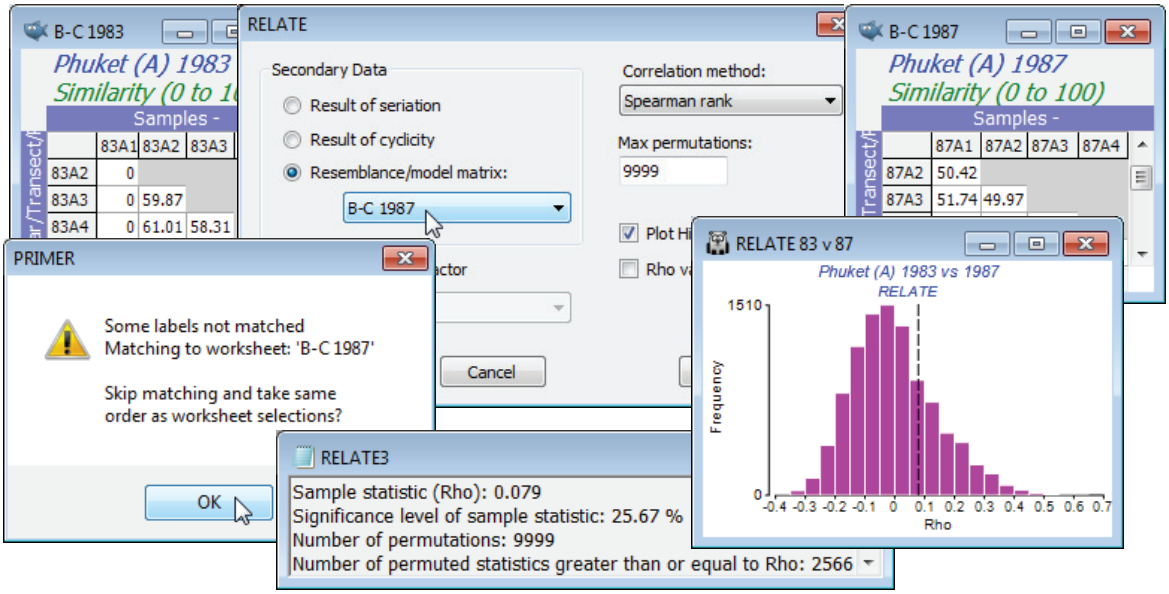
The serial change along the transect in 1983 has largely disappeared in 1987, with sedimentation impact from nearby dredging for a deep-water port. This is reflected in the RELATE tests shown above, with ρ declining from 0.651 to 0.193, e.g. on B-C 1983, **Analyse>RELATE>**(Secondary Data•Result of seriation) & (Max permutations: 9999), with defaults for the other choices, gives a histogram and results window with observed ρ = 0.651 greater than for any of the 9999 simulated values, so the null hypothesis of no seriation at all (ρ ≈ 0) is decisively rejected, p<0.01%. Note that the strong outlier has not wrecked this test, though it somewhat degrades the match to a model of equi-stepped change, as is seen by ρ rising to 0.75 if this first transect position is omitted. [Since we have not provided the factor Position when using the (•Result of seriation) option, the routine has to assume that samples are in the desired equi-stepped serial order – a different order, or a wish to fit unequal steps, perhaps by omission of an intermediate transect sample, must be handled by a Model Matrix.] The ρ = 0.193 for 1987 is more in the body of the null distribution however, and there is no clear evidence in the RELATE test for any serial structure (p≈7.5%). In this simple case, there is a very close link with the ordered, unreplicated 1-way ANOSIM test on factor Position (see Section 9), with R (not ρ) statistics of R^{Os} = 0.655 (p<0.01%) and 0.194 (p≈7.0%) for 1983 and 87.

v7

RELATE test on two biotic arrays

Given the breakdown of the serial gradient structure for 1987, is it now the case that the pattern of change down the transect has nothing at all in common with that for 1983? To answer that question requires a further run of RELATE, but of the two similarity sheets B-C 1983 and B-C 1987 against each other, rather than in comparison with a model matrix. With either as active window, say B-C 1983, take **Analyse>RELATE>**(Secondary Data•Resemblance/model matrix: B-C 1987). There will be a warning message indicating that the sample labels in the two sheets could not be matched. This issue was raised earlier, in Section 11. PRIMER typically takes label matching very seriously. When linking separate data sheets, as in **RELATE** or **BEST** (or the ABC plots of Section 16), the sample order need not be the same in the two matrices – provided it can find all the sample labels of the active matrix somewhere in the secondary sheet, the correct match will take place. However, it is here inconvenient to have to rename both sets of labels (currently 83A1, 83A2, … and 87A1, 87A2, …) to a common set (A1, A2, …), especially because the data were extracted from a larger sheet, where PRIMER expects the sample labels to be unique! So, this warning message provides an over-ride (take **OK**) which allows you to skip label matching, and RELATE will pair up the samples in the current order in both sheets. The option will not be offered if the two similarity matrices are not the same size. Instead you will get an error message *No labels matched. Cannot match labels, even relaxed*. The routine will then need to be run again, having selected the same number of samples in each, and it is your responsibility to make sure they are in the same order!

The results do indeed show that the assemblage patterns down the transect in the two years are totally unrelated. The observed match of only ρ = 0.079 is exceeded by about 2500 of the 9999 permutations under the null hypothesis (p<25%) – the null hypothesis (as always) being that there is absolutely no match in spatial pattern (ρ = 0). Omitting the outlier (Position 1) from both series, makes little difference to this conclusion, ρ now dropping still further to 0.016 (p<44%).



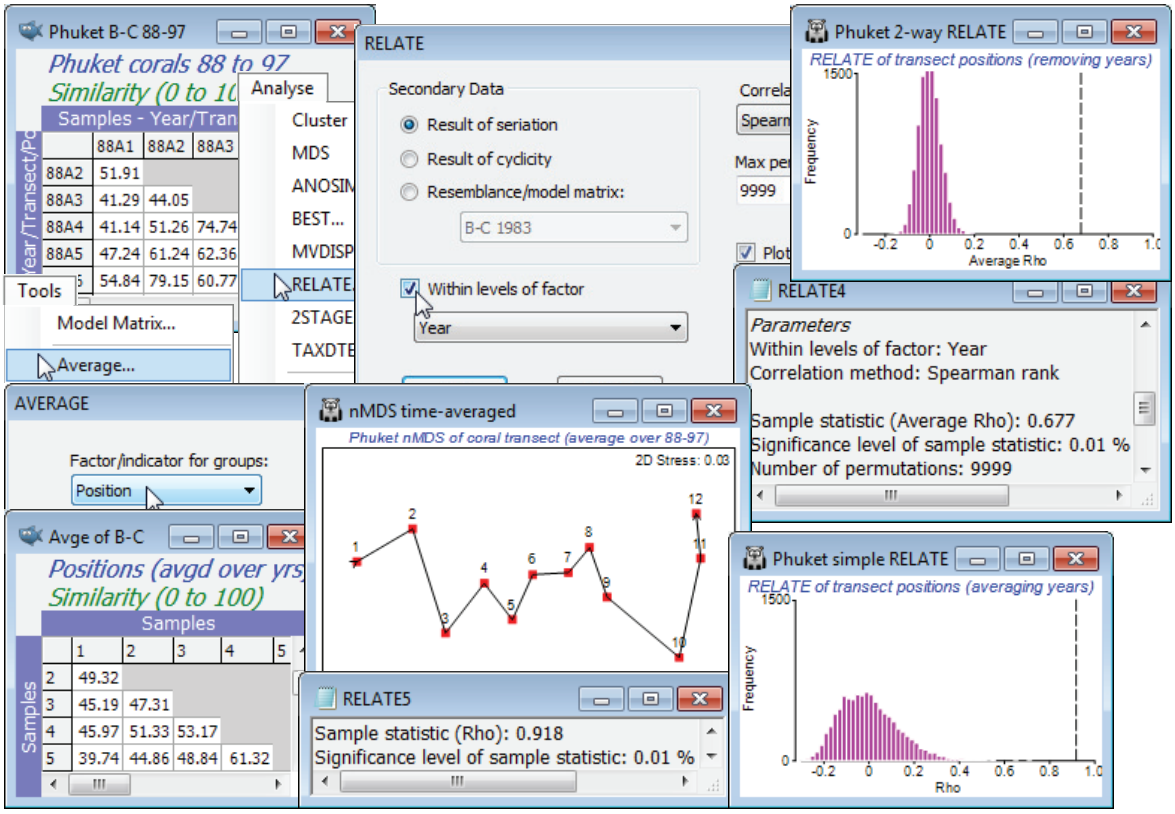
2-way RELATE for seriation

v7

A new feature in PRIMER 7 parallels that discussed for the BEST analysis of the previous section, namely a secondary factor is supplied, e.g. (✓Within levels of factor), which turns this into a *2-way RELATE test*. The matching statistic ρ – whether that is to simple seriation, simple cyclicity or a supplied resemblance/model matrix – is calculated only on samples within the levels of this secondary factor, and the ρ values averaged to give the overall test statistic. The permutations for the test are similarly constrained to be within the strata of this eliminated, secondary factor.

v7

In the current Phuket ws workspace, open (if necessary) the data file Phuket coral cover 88-97, and run the same root-transform and similarity as above, for these 12 transect positions × 7 years – a period with no known new stressors on the coral reef. Running RELATE on this similarity matrix, under the simple (•Result of seriation) model, and for (✓Within levels of factor Year), removes the inter-annual differences by calculating a simple trend statistic ρ across the transect positions, separately for each year, and then averaging those. (Note again that since we have not supplied the Position factor in setting up the test, the routine presumes that the samples for each year are in the desired serial order in the matrix). This is now a test statistic for the null hypothesis of no serial community change along the transect in any year, giving a large average ρ of 0.68, and obviously an overwhelmingly significant result ($p<0.01\%$). An alternative test would have been to average the samples over the years for each transect position (by **Tools>Average** using factor Position, on either the transformed data or the similarity matrix) and perform a simple seriation test on the 12 samples of the resulting matrix. As the *n*MDS plot for these averages shows, there is a very steady time-averaged gradient of change along the transect with RELATE statistic $\rho = 0.92$ ($p<0.01\%$). However, the histogram of the null distribution is seen to take values up to 0.3 or 0.4, in contrast with that for the 2-way RELATE test for which values larger than about 0.1 will be significant. In other words, 2-way RELATE is the more powerful test – by eliminating the year differences rather than averaging over them it has many more permutations and this could be important for testing very short runs of serial change (an averaged test with 5 transect positions has only (5!/2) = 60 distinct permutations, at best an ≈2% level test, and 4 positions is not viable, with 12 permutations).



v7

The above example was met in Section 9, under the 2-way ordered, unreplicated ANOSIM test, and there is again a very close affinity of the average ρ statistic with the average ANOSIM R^{Os}. In fact, there is no advantage here in using the 2-way seriation RELATE test – the equivalent ordered ANOSIM test is marginally preferable (see Chapter 6 in CiMC on ANOSIM for ordered factors). However, ordered ANOSIM is constrained to the simple serial model, whereas 2-way RELATE

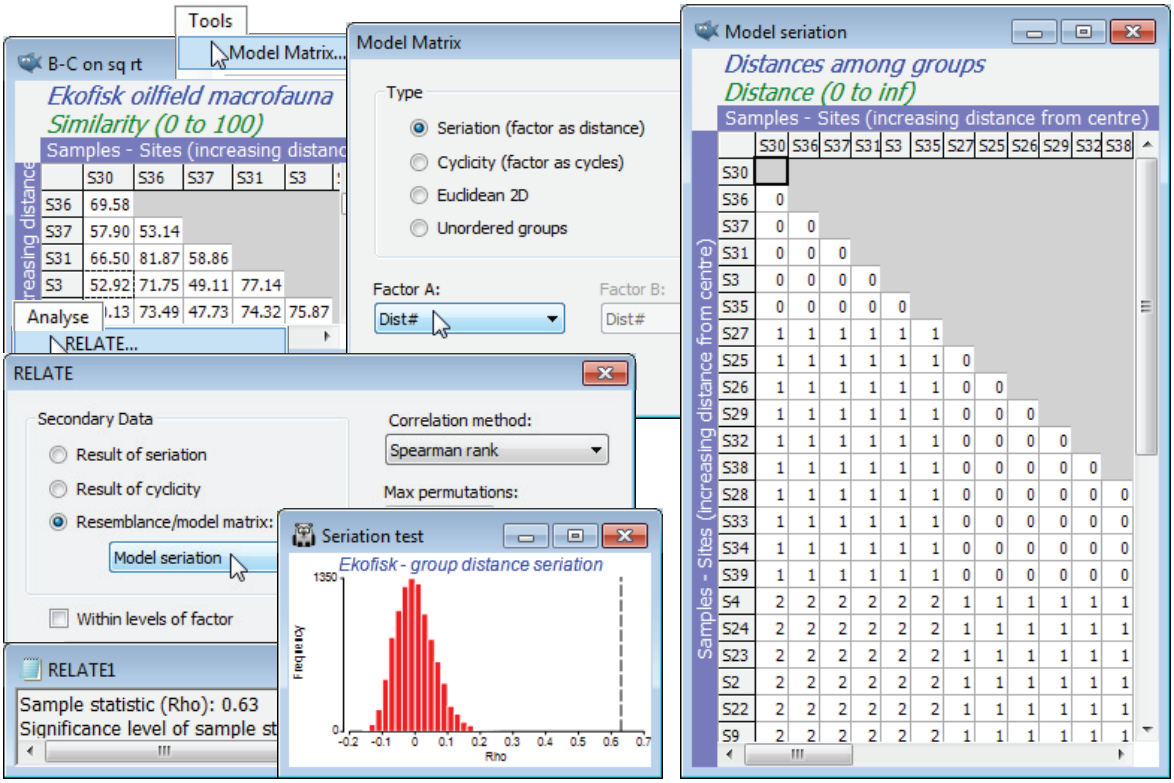
comes into its own, later, when we move to other model matrices, e.g. seriation for a time series where the times are not equally spaced, and we wish to allow for this in computing the statistic (though in most cases that will make very little difference because of the rank nature of the tests) or, more importantly, when the model is not serial but cyclic, or based on a supplied resemblance matrix from abiotic variables, perhaps. This takes us back to the 2-way BEST construction of the previous section – matching to an environmental resemblance matrix, having removed a categorical factor. The difference of 2-way BEST from 2-way RELATE, of course, is that between any global BEST test and an equivalent RELATE test – the former allows for the selection bias in repeating abiotic variable choices until the best match is found, whereas the latter assumes a single fixed set.

v7

Save the workspace Phuket ws, which will be returned to later in the context of 2nd stage analysis.

Seriation with replication

Return to the macrofaunal data set from the Ekofisk oilfield, with workspace Ekofisk ws last saved in C:\ Examples v7\Ekofisk macrofauna in Section 9, following an ordered 1-way ANOSIM test (with replication) on the similarity matrix B-C on sq rt from data Ekofisk macrofauna counts. This used factor Dist#, which is the numeric form of the four groups of sites at different distances from the oilfield, ~logarithmically spaced (1≡D:<250m; 2≡C: 250m-1km; 3≡B: 1-3.5km; 4≡A:>3.5km). The rationale for an ordered test here was discussed in Section 9 (and Somerfield PJ, Clarke KR, Olsgard F 2002, *J Anim Ecol* 71:581-593), namely the improved power but more limited generality in testing the null H_0: *no differences* against an ordered alternative H_1: $A \rightarrow B \rightarrow C \rightarrow D$, rather than the unordered alternative H_1: *A, B, C, D differ* (in ways unspecified). Those authors, and previous versions of PRIMER, did not use the generalised (ordered) ANOSIM statistic – which is new to PRIMER 7 – but used the analogous RELATE statistic ρ between the biotic resemblances and a model matrix for *seriation with replication*. This is a model matrix which **Analyse>RELATE** does not handle internally in the (•Result of seriation) option – that is restricted to simple seriation with no replication – but which can be simply constructed from the active matrix B-C on sq rt, using **Tools>Model Matrix>**(Type•Seriation (factor as distance)) & (Factor A: Dist#). [The factor Dist, splitting the sites into alphabetic levels D, C, B, A, will not work here because distances cannot be calculated between names]. A model matrix is generated – rename this Model seriation – having blocks of 0's down the diagonal (sites within a distance group are considered 0 distance apart), then off-diagonal blocks of 1's then 2's then 3's (sites in groups 1 and 2 are 1 unit apart, in groups 1 and 3 are 2 units apart etc). Again with B-C on sq rt active, run **Analyse>RELATE>**(Secondary data• Resemblance/model matrix: Model seriation), giving ρ = 0.63 (p<0.01%), providing clear evidence of group differences, with large ρ confirming the strongly ordered gradient away from the oilfield.

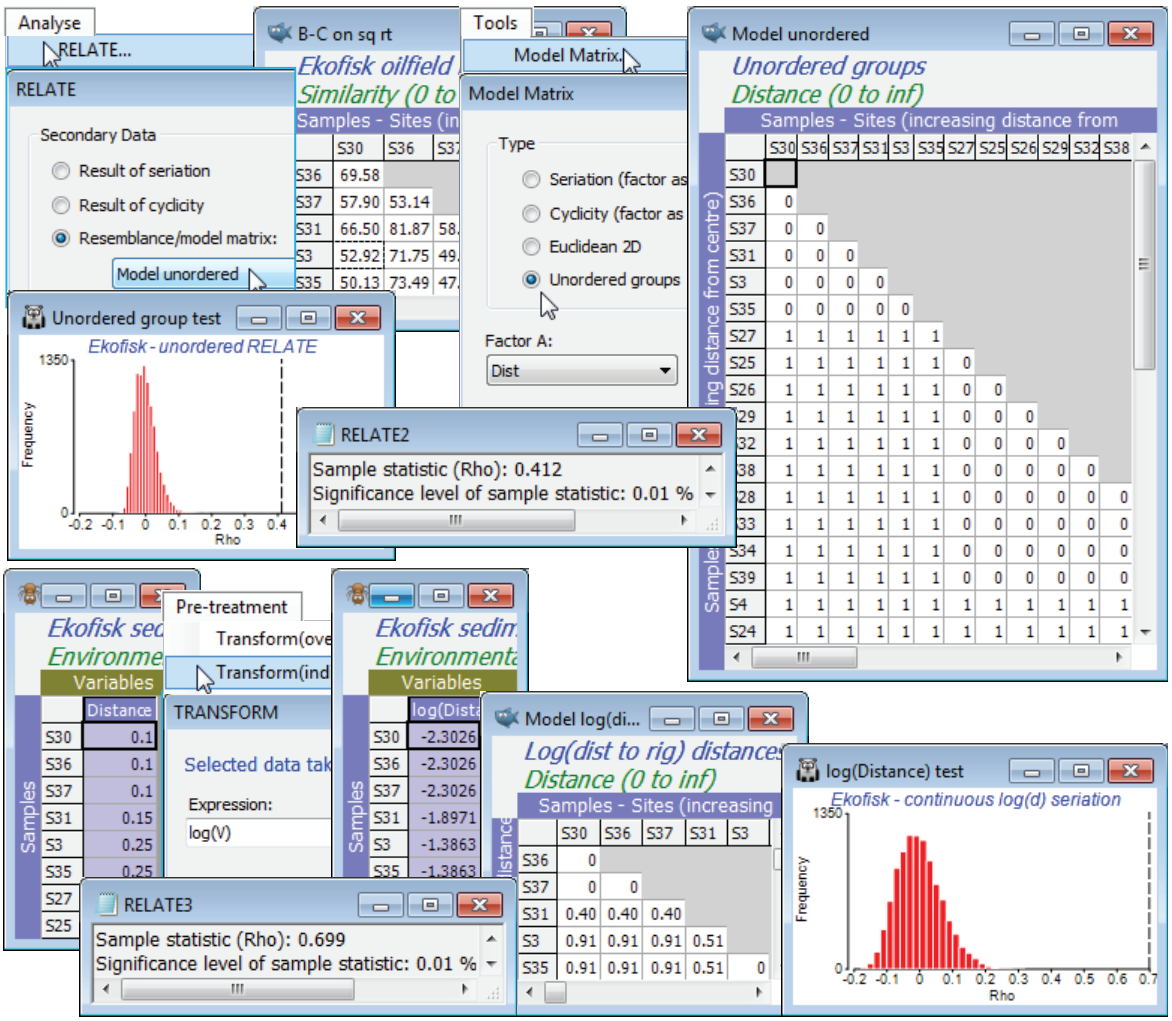


Other Model Matrix options

The conclusion is, of course, consistent with the different, but closely-related, ordered ANOSIM statistic R^{Oc} = 0.67 (p<0.01%), calculated in Section 9. It is relatively easy to show algebraically that the unordered ANOSIM statistic – here R = 0.55, with p<0.01% again – is exactly equivalent (see Chapter 6 of CiMC) to a RELATE ρ test with model matrix having 0's in the diagonal blocks (samples within the same group) and 1's elsewhere (samples in different groups), i.e. all groups are considered equally different from each other. Such a model matrix can be constructed by **Tools> Model Matrix>**(Type•Unordered groups) & (Factor A: Dist) – or (Factor A: Dist#), since it no longer matters whether an unordered factor is supplied as numeric or alphabetic. This test returns ρ = 0.41, again strongly significant, naturally, but much lower than the seriation statistic ρ = 0.63.

v7

So far we have not seen anything that could not have been slightly better carried out with ordered and unordered ANOSIM tests (in the sense that the statistic upper limit of +1 is attainable for R^O but not for ρ – see Chapter 6 of CiMC – and because ANOSIM will allow pairwise comparisons among the groups). But another possible model here which we would like a comparison with – one which can only be handled by RELATE – is to ignore the arbitrary distance group structure and RELATE the biotic similarities to the distance matrix calculated from (log-transformed) distances of each site to the oilfield centre. (The log transform reverses an exponentially decreasing dilution curve of contaminant concentrations with distance). The raw distances are in the first column in the abiotic sheet Ekofisk environment, the *Distance* variable (you may need to **Select>All** to find it!). Highlight and select just this column, log transform it by **Pre-treatment>Transform(individual)>** (Expression: log(V)) and **Analyse>Resemblance>**(Measure•Euclidean distance), renaming the result as Model log(distance) and inputting it as the secondary matrix to an **Analyse>RELATE** on B-C on sq rt. The result is again significant, naturally, but arguably demonstrates an even stronger gradient of assemblage change with this model of continuous (logged) distance from the oilfield, ρ = 0.70 (p<0.01%). This model matrix could also have been created by copying the log(Distance) entries into a factor log(D) under the B-C on sq rt resemblance sheet and running **Tools>Model Matrix>**(Type•Seriation (factor as distance)) & (Factor A: log(D)). Save and close Ekofisk ws.



Expanding an (abiotic) data matrix

A RELATE test could equally well have been carried out between the Ekofisk community pattern and a matching (abiotic) resemblance matrix computed not from the surrogate for increased impact – the nearness of the sites to the oil-field centre – but from a set of contaminant levels themselves, as measured at each site. (For the tests of this section, we assume that this set is fixed – we are not allowing selection of a subset of contaminant variables which appears to best match the observed community pattern, i.e. the BEST(Bio-Env) procedure of the previous section. RELATE tests do not allow for this selection bias). All that is necessary for a simple RELATE test of community to a fixed environmental variable suite is that we have one-to-one matching of the abiotic data to each community sample. The active sheet for **Analyse>RELATE** would logically be the biological resemblance coefficient and the (Secondary data•Resemblance/model matrix) would typically be Euclidean distance on a selectively transformed then normalised abiotic data matrix (though the test would be the same if the matrices were the same size and entered in the opposite order). But where the community data consists, for example, of replicate samples at a number of sites, and the abiotic matrix consists of a single value for each of the suite of variables (which may itself be an average over replicate abiotic measurements, but not matched to the community replicates) then the abiotic matrix needs to be expanded to the same dimensions as the biological matrix, and its entries repeated appropriately. This is achieved by the **Tools>Expand Samples** routine operating on the active matrix of the abiotic data. It is not cheating – at least, not necessarily! It depends on what is then done with the expanded matrix. If we pretend that the repeated readings are independently measured – by running an ANOSIM test on them for example – then of course we are heading for trouble. But in this context the requirement is an expansion of the *Seriation with replication* test of the previous page – we want to test the null hypothesis that there are no differences among sites against the specific alternative that there are such differences and that they are determined by the environmental structure among sites (in statistical parlance we *condition* on this, so the situation becomes no different than if we were testing against a design structure, e.g. seriation or treatment levels). So, the test is no longer of seriation with replication but of a more complex environmental relationship among the sites, but it has the same characterising feature that the resulting RELATE ρ value will capture both whether the sites differ at all and whether they do so in a way that matches the (multivariate) abiotic relationships among sites. A high ρ can only be obtained if both are true. An alternative would be to average up the community replicates to the site level and carry out a simple RELATE test to the abiotic data at that level. However, this might have very little power if there are few sites and it misses the important comparison of whether ρ for this specific alternative is greater than ρ for the unordered test (the 1-way ANOSIM-type model matrix of 0's and 1's).

v7

Expanded RELATE test (Exe nematodes)

As an example of **Tools>Expand Samples** on a data matrix (or **Tools>Expand** on a resemblance matrix, since the expansion can be equally well achieved either before or after the computation of Euclidean distance in this situation) we shall use the Exe nematode study and the form of the data met in Section 9, in which the 19 sites from different environmental conditions around the Exe estuary were sampled 6 times through one year (with just 6 missing samples spread over several sites, i.e. 108 meiofaunal core samples in total). The biotic matrix is Exe nematodes bi-monthly in C:\Examples v7\Exe nematodes, comprising abundances of 182 species (its time-averaged form was used extensively in Section 8). Also open the abiotic data, Exe environment, which we have not encountered here but which is used as a motivating example for the BEST routine in Chapter 11 of CiMC. It consists of 6 sediment-based environmental variables, postulated to be structuring the communities of free-living nematodes, and recorded as relevant to each site over the full year of sampling: median particle diameter, depth of the water table, depth of the blackened H_2S (anoxic) layer, height up the shore (this was an intertidal study), % organics and the interstitial salinity. The environmental data therefore has only 19 samples, which are labelled with the site numbers (1-19). Importantly for the **Expand** routine, those labels need to be exactly the same as the levels for the site factor which is defined for the 108 samples of the Exe nematodes bi-monthly biotic matrix.

v7

v7

The biotic samples do have a time (i.e. seasonal) structure, in that they are all collected bi-monthly – factor time, with levels A, B, C, D, E, F in common for each site. The time factor will be ignored, however, for the purpose of this illustration and the (up to) 6 values used as replicates for each site. This is not unreasonable, since they will represent both the spatial and temporal variability at that site through the year (providing a conservative estimate of the true residual variability) and it was seen earlier – in Section 9 for sites 12-19 but true also for all sites – that 2-way crossed ANOSIM (without replication) fails to find significant evidence of a seasonal effect at all.

The Exe nematodes bi-monthly biotic matrix requires fourth-root transformation before the usual similarity calculation (resemblance B-C 4rt), and the *n*MDS plot for all 108 samples, with symbols as the site and labels removed (from **Graph>Sample Labels & Symbols**), shows clear differences among sites. The unordered RELATE test – equivalent to unordered 1-way ANOSIM but giving a ρ test statistic which we can compare with the expanded abiotic test – is obtained by running on the active B-C 4rt sheet: **Tools>Model Matrix>**(Type•Unordered groups) & (Factor A: site) to give the Unordered model. Then, again on B-C 4rt, **Analyse>RELATE>**(Secondary data•Resemblance/ model matrix: Unordered model) gives a Spearman rank statistic of ρ = 0.33 (p<0.01%) – though highly significantly different from zero (and thus confirming site differences), ρ is not large.

Expand Samples or Expand resemblances

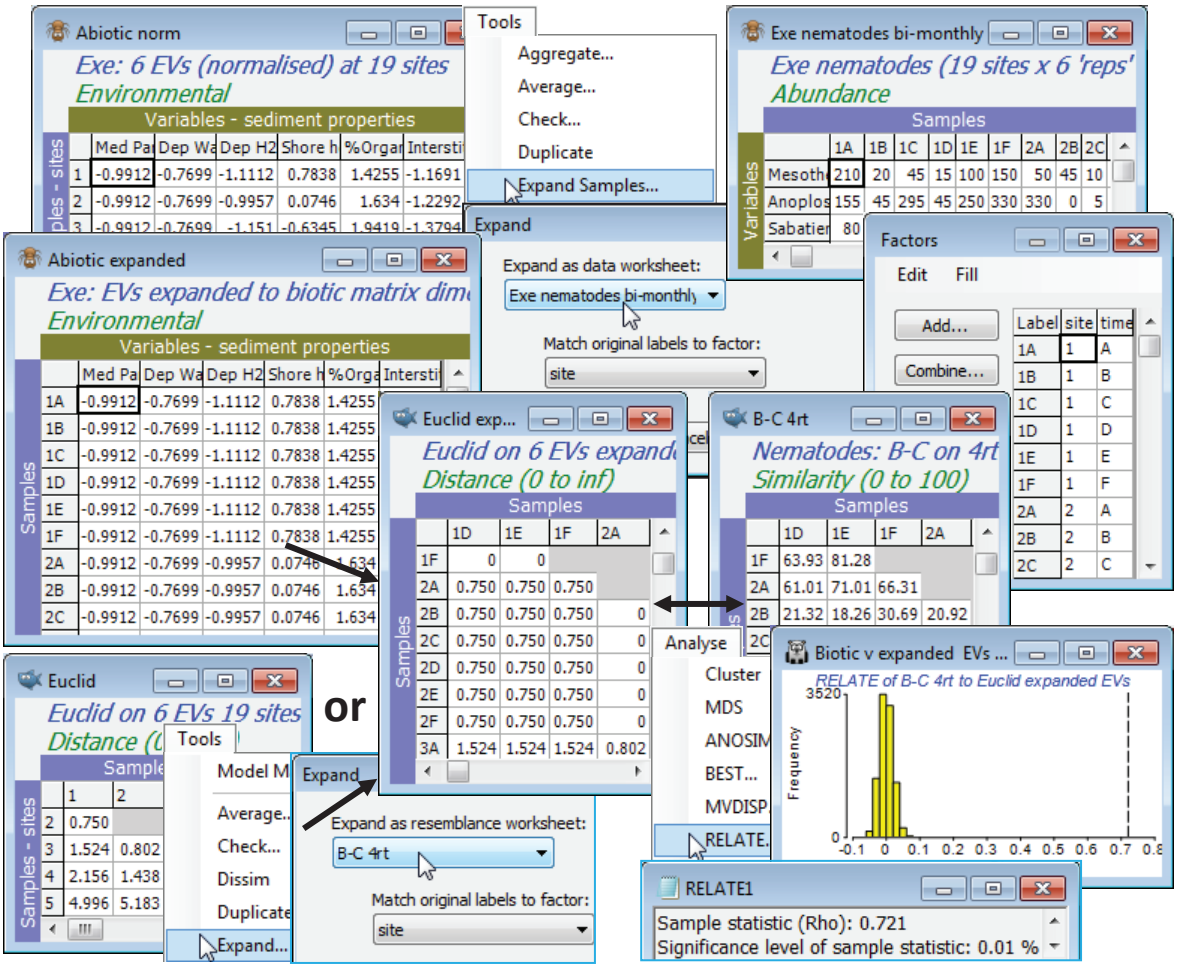
v7

The Exe environment matrix does not seem (from **Plots>Draftsman Plot** or **Histogram Plot**) to contain notable outliers and can safely be used without transformation of individual variables. It does however need **Pre-treatment>Normalise Variables** – rename it Abiotic norm. To expand this data matrix to the dimensions of 108 samples × 6 variables, with Abiotic norm as the active sheet take **Tools>Expand Samples>**(Expand as data worksheet: Exe nematodes bi-monthly) & (Match original labels to factor: site). The fourth-root form of the biotic data matrix could equally well have been used in place of the original nematode sheet – what is needed from it is the size of expanded matrix needed, and the structure of samples over the sites, from the factor site whose levels 1, …, 19 are matched up with the labels 1, ..., 19 of the normalised environmental sheet. The expanded environmental matrix is then entered to a Euclidean distance resemblance calculation, to give Euclid expanded. The same construction can be achieved by first taking Euclidean distance on the Abiotic norm data matrix, to give a resemblance matrix renamed Euclid, and then entering this

v7

as the active matrix in **Tools>Expand>**(Expand as resemblance worksheet: B-C 4rt) & (Match original labels to factor: site) to obtain exactly the same model (abiotic) matrix Euclid expanded.

Now with active sheet B-C 4rt a further run of **Analyse>RELATE>**(Secondary data•Resemblance /model matrix: Euclid expanded) gives a much larger ρ of 0.72 (highly significant, of course, at p<0.01% for the 9999 permutations of this run), indicating the very good fit of the individual bi-monthly samples to the alternative model of sites differences, structured by these abiotic variables.

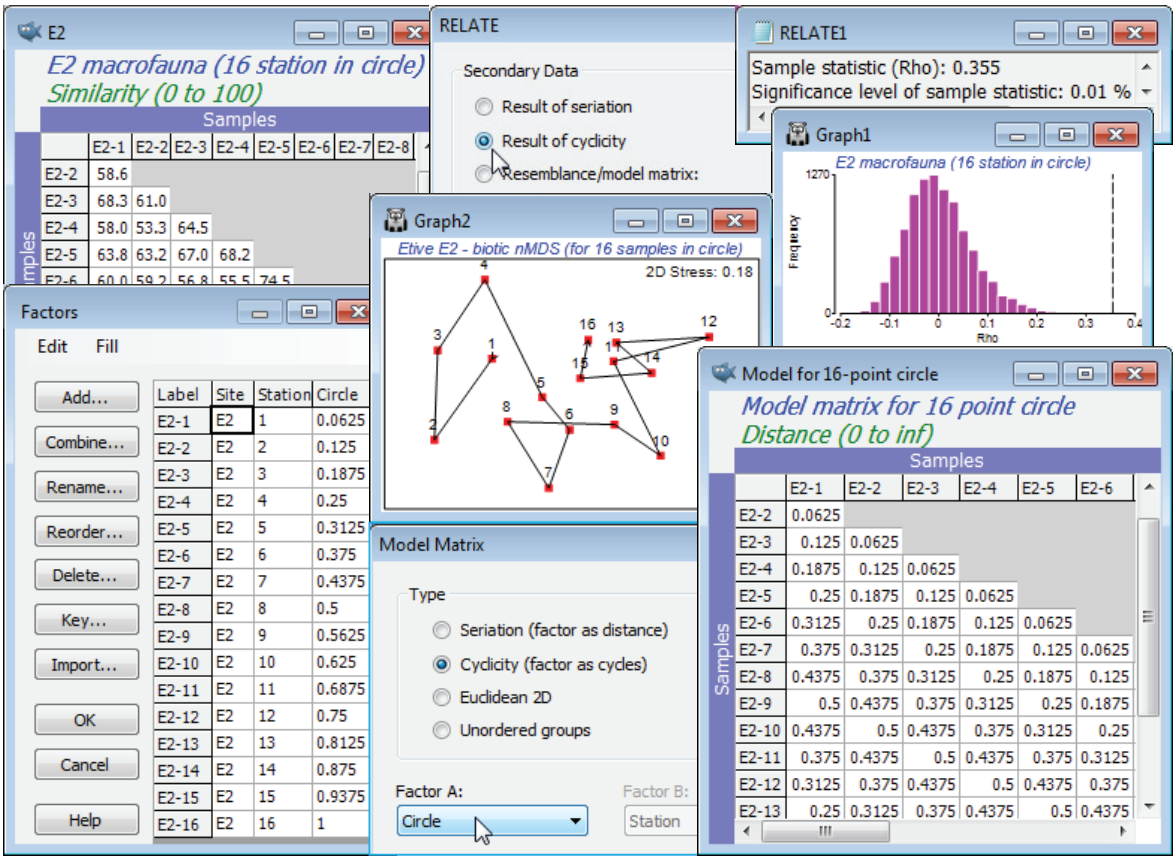


Model matrix for 2D Euclidean

The other two **Model Matrix** options are (Type•Cyclicity (factor as cycles)) and (Type•Euclidean 2D). The latter simply calculates, for example, distance between samples in a geographic layout when the *x, y* co-ordinates of the sample points are not held in a separate (environment-type) data sheet but as numeric factors in the biotic data. The corresponding model for sample locations in a 1D layout, given by a single factor, is just the (Type•Seriation) option, or equivalently, set up a Factor B with the same level (e.g. 1) for all samples and take the (Type•Euclidean 2D) option.

Cyclicity (Sea-loch macrofauna)

The (Type•Cyclicity (factor as cycles)) option in **Model Matrix** needs a numeric factor over the range (0, 1), representing the distances round a circle, where 0 and 1 are at the same point (or think of these as the angles at which those points are set, ranging over 0 to 1, not 0 to 360). The obvious examples of such data are in a time-series over a full seasonal cycle (see shortly), or a diel or tidal cycle, but we shall start with an unusual spatial example from studies described by Gage JD 1972 *Mar Biol* 14:281-297 (and analysed in a multivariate way by Somerfield PJ & Gage JD 2000 *Mar Biol* 136:1133-1145), of subtidal macrobenthos in Scottish sea-lochs. The subset of these samples used here is from three sites in Loch Etive, at each of which 16 samples (factor Stations, 1-16) were taken approximately around the circumference of a 100m diameter circle at equal spacing. Over all three places (factor Site, E2, E6, E24) counts were made of a total of 186 species, in data file Etive macrofauna counts, directory C:\Examples v7\Sea-loch macrofauna. Open the data and select, for the moment, just the 16 samples from Site E2, by **Select>Samples>**(•Factor levels)>(Factor name: Site)>**Levels>**(Include: E2) & (Available: E6 & E24). On fourth-root transform and Bray-Curtis similarities, run the *n*MDS ordination and test the null hypothesis that there are no differences in communities at these 16 stations against the alternative of a circular structure (their spatial layout) with **Analyse>RELATE>**(Secondary data•Result of cyclicity). For simple cyclicity such as this, with equal spacing, no replication at each point, and an assumption that the stations are in correct order round the circle, the routine creates the model internally, and an explicit construction of the model matrix is not needed. However, it is instructive to create the model externally, from active sheet of the biotic similarities of the 16 samples, using **Tools>Model Matrix>**(Type•Cyclicity). The supplied (Factor A:) is not the Stations levels 1-16, but these numbers divided by 16, held in the factor Circle – in general the points, e.g. times, may not be equally spaced and the routine must be told how the start and end of the sequence relate to each other, hence the restriction to (0,1). The same RELATE test now results from using this cyclic model under (•Resemblance/model matrix: Model for 16-point circle), giving weak but still significant cyclicity (ρ = 0.355, p<0.01%).



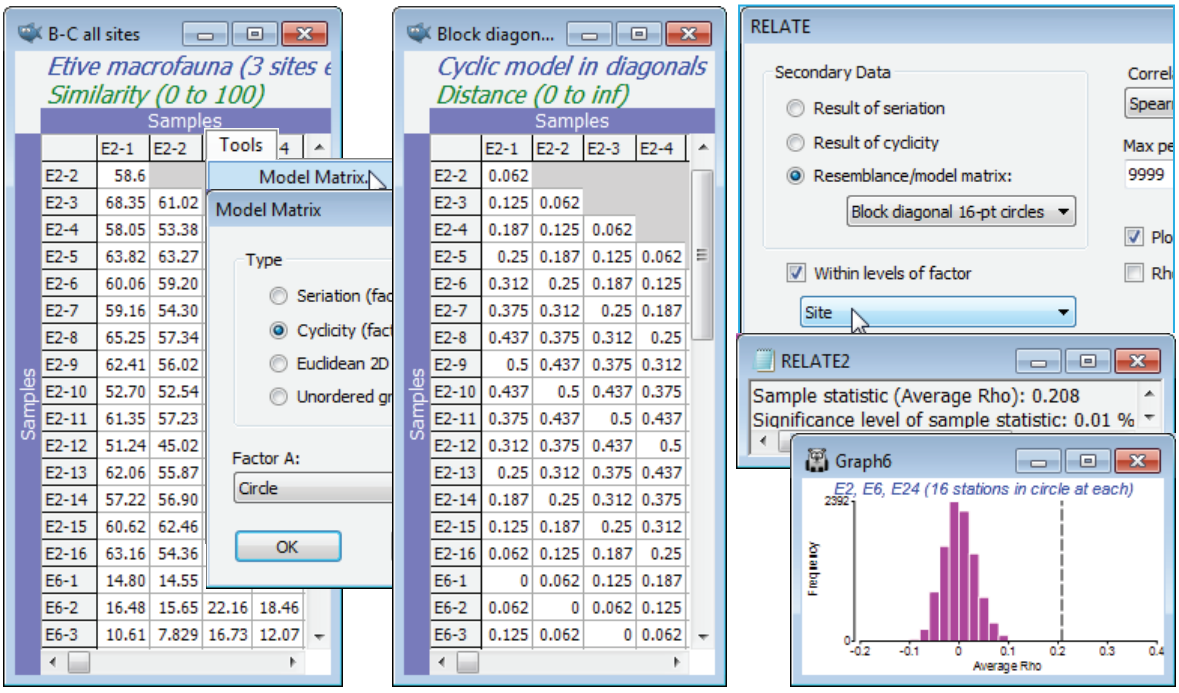
2-way RELATE for cyclicity

v7

A 2-way RELATE version of the above test where there are no replicates, and the cyclic factor under test is actually nested within a 'nuisance' factor whose effect we want to remove, is given by reverting to the full data sheet for the Loch Etive macrofauna samples: **Select>All** and recompute the fourth-root transform and Bray-Curtis similarities, as B-C all sites, which now has 16 samples in a circle for each of the three sites E2, E6 and E24. Testing for a match with the circular spatial layout of stations, simultaneously at all three sites, whilst eliminating the inevitable differences in community composition for these three locations using 2-way RELATE, should give a still stronger test of the null hypothesis of no community differences within sites against this specific alternative.

v7

As before, the model matrix is constructed by **Tools>Model Matrix>**(Type•Cyclicity) & (Factor A: Circle), run on B-C all sites, giving a model distance matrix (rename it Block diagonal 16-pt circles) in which only the block diagonals of the stations within sites will be sensible in this nested case (and which is all that RELATE uses) because, for example, station 1 at E2 and station 1 at E6 have nothing in common. On B-C all sites, **Analyse>RELATE>**(Secondary data•Resemblance/model matrix: Block diagonal 16-pt circles) & (✓Within levels of factor Site) gives an averaged ρ statistic across the three sites of 0.21, still strongly significant – note the tighter spread of the null histogram (c.f. Graph1 above) because of the simultaneous testing. The lower value than for the E2 test alone suggests weaker effects at E6 and E24, which is seen in separate (1-way) cyclic tests.



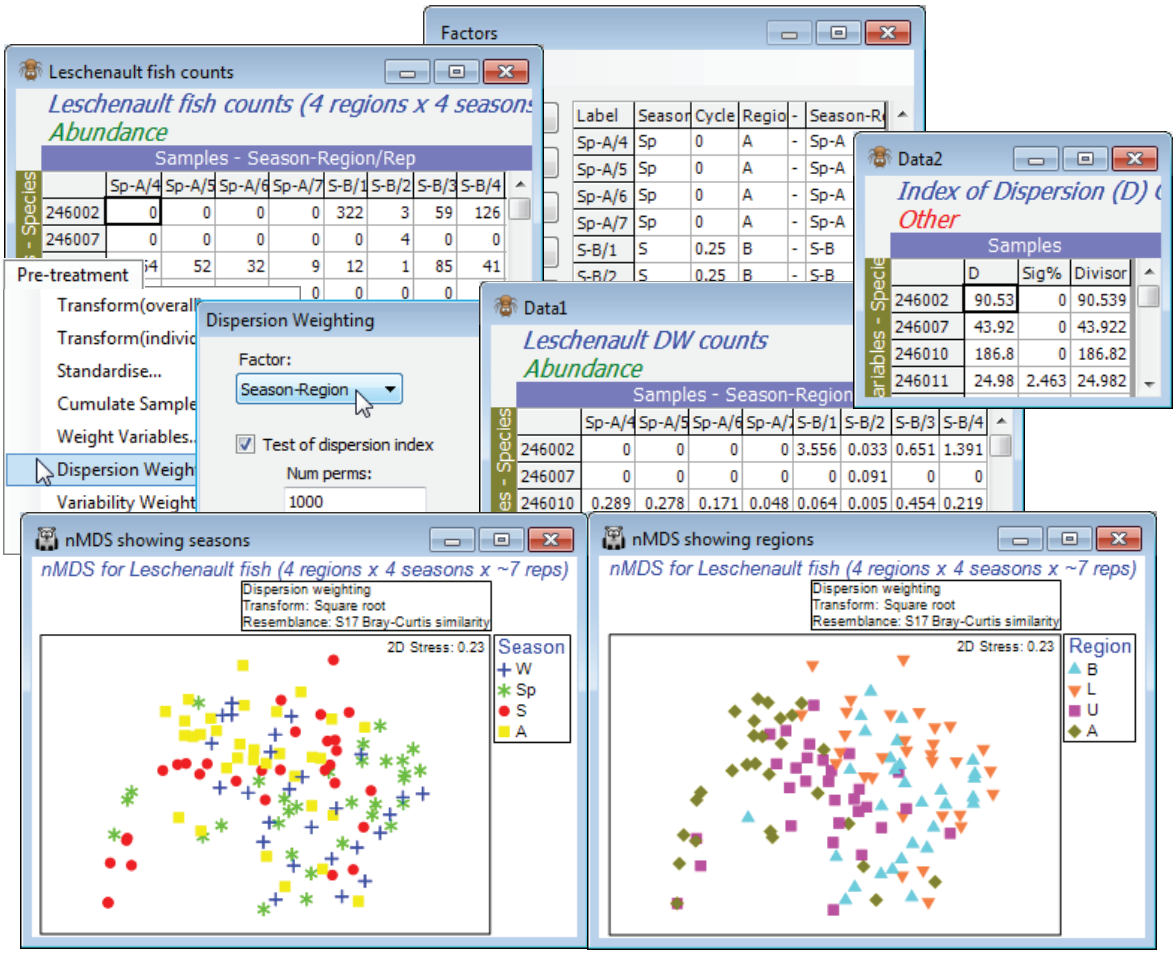
v7

More usually, the cyclic factor under test (often time) is crossed with a second factor (often space), whose effect we want to eliminate for our time test. The 2-way RELATE test structure is the same as for the above nested case however, and a more typical example is now given of a cyclic four-seasons series recorded for several regions, with the added complexity of replication within each of the cells of this 2-way layout. Though the structure is that of a 2-way crossed ANOSIM, this case is not covered by running an ordered ANOSIM test because, of course, the time factor is cyclic and not serial – an appropriate model matrix therefore needs to be created as an input to RELATE.

(Leschenault estuarine fish, W Australia)

Veale L *et al* 2014 *J Fish Biol* 85: 1320-1354 describe trawl sampling for nearshore estuarine fish in the Leschenault estuary of Western Australia, over 4 regions (B - Basal, L - Lower, U - Upper, A - Apex of the estuary) and 4 seasons (Sp - Spring, S - Summer, A - Autumn and W - Winter). The data set used for this illustration has been somewhat simplified and consists of 6-8 replicate 21.5m seine net samples reflecting both inter-annual and spatial variation within each of the 16 region×season combinations. Due to the location of freshwater inputs and restricted exchange with the ocean, the estuary has a salinity gradient which increases from the basal (mouth) through lower and upper regions to the estuary apex. Counts are given of 43 fish species (with numerical ID), file Leschenault fish counts in C:\Examples v7\Leschenault fish. Close the above workspace and open this file, with factors Season and its (0, 1) numeric form Cycle (0, 0.25, 0.5, 0.75), then Region.

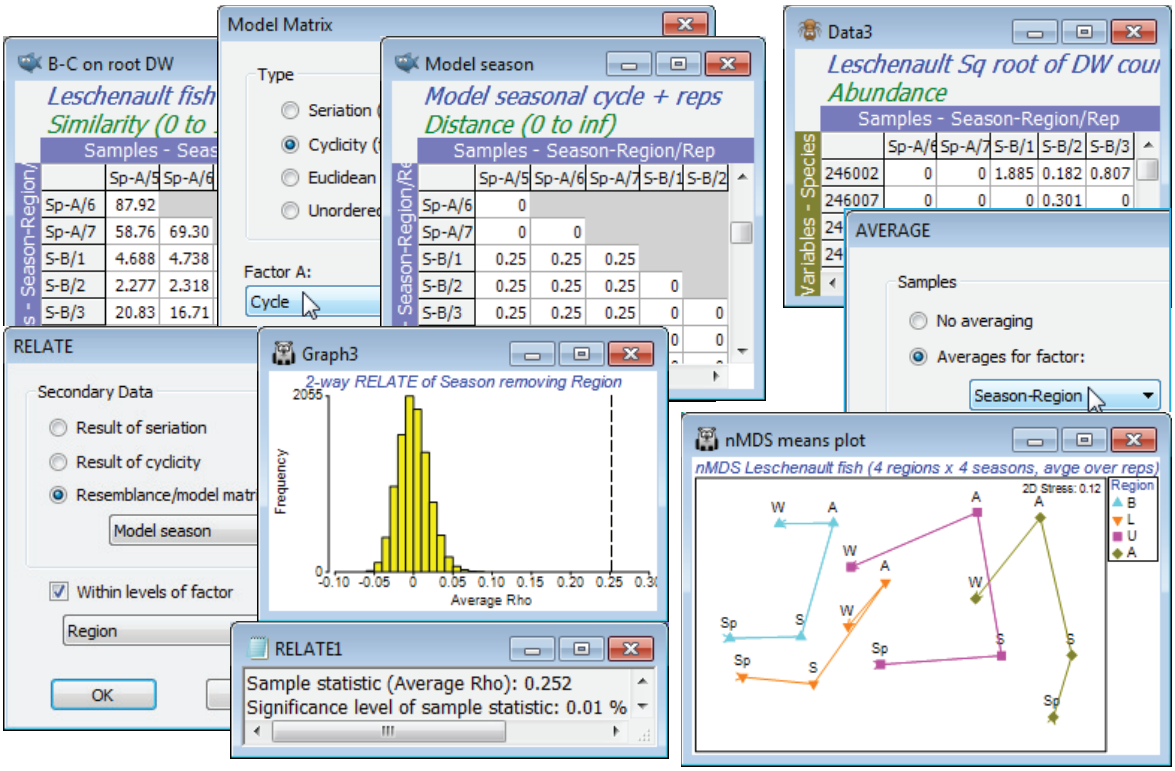
As often with fish data, over-dispersion of counts (shoaling) can be substantial for some species, their erratic counts over replicates giving them too much weight in a community assessment, and Clarke KR, Tweedley JR, Valesini FJ 2014 *J Mar Biol Ass UK* 94: 1-16 show that a good strategy for such fish data is often pre-treatment by Dispersion Weighting (Section 4 and Chapter 9, CiMC) followed by mild transformation (square root). So, take **Edit>Factors>Combine>**(Include: Season & - & Region), where - is just a hyphen separator in all rows, to create a new factor Season-Region whose levels identify the groups of replicates from the 16 conditions. Use this in **Pre-treatment> Dispersion Weighting>**(Factor: Season-Region) & (✓Test of dispersion index) &(✓Stats to worksheet), and the latter sheet shows that counts of some species are, indeed, heavily downweighted by
v7 an index of dispersion D of up to nearly 200. However, **Plots>Shade Plot** (Section 4) or **Wizards> Matrix display** (Section 10) on the dispersion-weighted data still show that the contributions to the resemblance matrix will come from relatively few of the species, so take a further **Pre-treatment> Transformation(overall)>**(Transformation: Square root). Now calculate Bray-Curtis similarity on this full set of 119 samples (B-C on root DW), and an *n*MDS ordination with symbols for Region, and duplicated with symbols for Season, show a great deal of replicate variability and consequent high stress, but also some evidence for effects of both factors.



Unordered two-way ANOSIM with factors Region and Season is perfectly viable and will provide
v7 pairwise comparisons, though there is a good case for 2-way ANOSIM with an ordered Region factor, because of the salinity gradient and the geographical ordering of the regions B, L, U and A (note that a numeric factor would need to be created to capture this order). Such a serial order is not appropriate for Season, however, with the cyclic relationship of its levels (the factor Cycle, with Sp = 0, S = 0.25, A = 0.5, W = 0.75; though Sp could equally well have been coded 1, of course). The optimum test of Season therefore creates a model structure from B-C on root DW with **Tools>**
v7 **Model Matrix>**(Type•Cyclicity) & (Factor A: Cycle) to give Model season and tests it by 2-way RELATE on the active sheet B-C on root DW with **RELATE>**(Secondary data•Resemblance/ model matrix: Model season) & (✓Within levels of factor: Region). The resulting match to a cyclic seasonal pattern in each region – under the 2-way model, separately calculated then averaged to give $\rho = 0.25$ – is low but this simply reflects the high replicate variability and therefore the strong

overlap of the communities in the different seasons for the same region. Importantly, this value is highly significantly different from zero, as the histogram shows (p<0.01% since 9999 permutations were again used). This certainly justifies an *n*MDS means plot, averaging the replicates for the 16 conditions (4 seasons × 4 regions). As we have seen, there are several possible ways to do this – averaging the replicates of the original counts, or the dispersion weighted and transformed data, or the similarities (or, in PERMANOVA+, using distances among centroids in the high-d PCO space). Here, take the second method, **Tools>Average>**(Samples•Averages for factor: Season-Region) on the transformed DW data matrix, then recalculate the Bray-Curtis similarities and the *n*MDS, on which display symbols as Region as labels as Season using **Samp. Labels & Symbols**, and overlay split trajectories using **Special>Overlays>**(✓Overlay trajectory:Cycle)>(✓Split trajectory:Region). Both the consistent community change up the estuary (B,L,U,A) and the matching seasonal cycles are evident. (Lines logically joining W and Sp, as in Fig. 15.12 of CiMC, can be added by copying and pasting the plot into Powerpoint, or similar software, where it can be ungrouped to Microsoft drawing objects and manipulated as vector graphics). Close the Leschenault workspace.

v7



Rationale for 2nd stage MDS

As seen above, the ρ statistic, which rank correlates the elements of two similarity matrices, can provide a very useful and succinct summary of the extent of agreement between two ordinations (or, to be more precise, of agreement in the high-dimensional multivariate data underlying these low-dimensional plots). Often, many such pairwise comparisons are made; for example, a single set of data may first be aggregated to a range of taxonomic levels (species, genus, family, …), then analysed under a range of pre-treatments: standardisations (none, by species or samples, and by maximum or total); other taxon weightings (e.g. dispersion weighting); then transformations (none, square root, 4th root, log, pres/abs), etc. Many ordination plots result and it is reasonable to ask how much the multivariate pattern changes as a result of these various decisions. What are the important choices? Does it matter whether the data are only identified to family rather than species level, or is the difference this makes completely dwarfed by the changes resulting from choosing to look at common to mid-abundance species (none or square root transform) or concentrating more on the less-common species (4th root or presence/absence)? Or is it the choice of a resemblance coefficient (from the 40 or so in Section 5) that really dictates the conclusions? It can be difficult, and arbitrary, to assess this just by looking at the range of different ordinations produced, though at least we can exploit the ρ statistic to give quantification of the agreement in multivariate pattern for any pair of choices. But when there are many choices, even a set of ρ values between pairs does not become a succinct enough description (considering only two types of choice, there are 20 different ordinations from 5 transformations and 4 taxonomic levels, thus 190 ρ values between them!).

The key step here is to realise that ρ itself can be regarded as a similarity measure, taking values near 1 if two multivariate patterns are highly similar and near zero if they bear no relation to each other. So, the triangular matrix of ρ coefficients between all pairs of ordinations can be entered into the MDS routine, to obtain what PRIMER calls a *2nd stage MDS plot* (an MDS of MDS's, if you like!). The ρ coefficient is not a distance-like measure (it can take small negative values and has a fixed upper limit) so it is unlikely to be turned into an ordination distance by a straight line through the origin on a Shepard plot, so again *n*MDS rather than *m*MDS seems appropriate This is based on the rank orders of the ρ values, therefore catering naturally with the potential for small negative ρ values – these just become patterns that are even less like each other than random re-arrangements, and in practice large negative values are not observed. The resulting second-stage *n*MDS plot thus gives a succinct summary in a 2-d picture, often with small stress, of the relationship between the multivariate sample patterns under the various choices. The **2STAGE** idea was introduced in this context by Somerfield PJ & Clarke KR 1995 *Mar Ecol Prog Ser* 127:113-119 and further explored by Olsgard F, Somerfield PJ, Carr MR 1997 & 1998 *Mar Ecol Prog Ser* 149: 173-181 & 172: 25-26, and is also covered extensively in Chapter 16 of CiMC, including the examples below.

v7

Aggregation & transforms (Morlaix macrofauna)

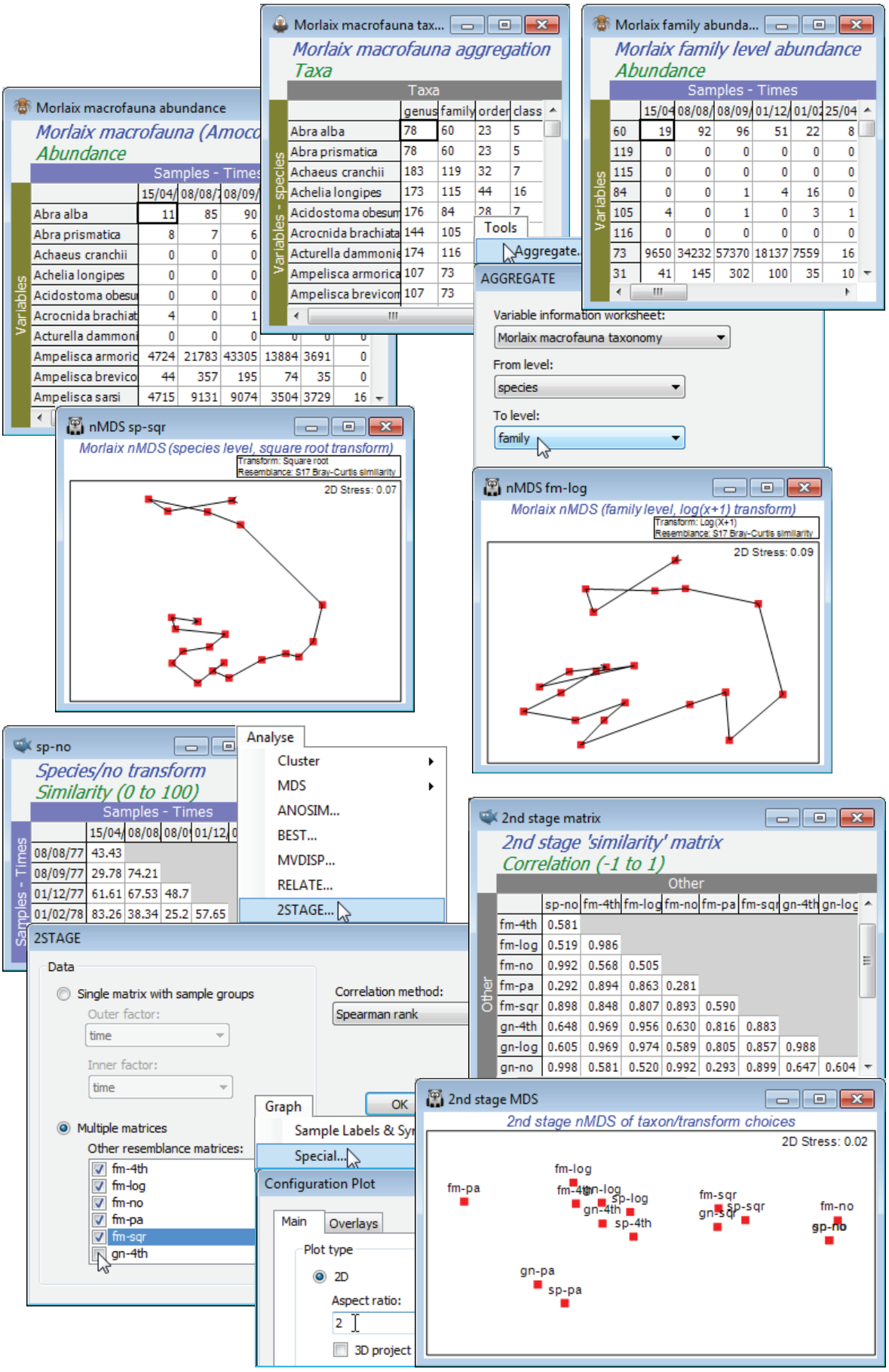
Chapter 10 of CiMC gives several examples of aggregating species matrices to higher taxa – using the **Tools>Aggregate** routine – and the effect this has on the resulting multivariate (and univariate) analyses. We shall illustrate this with the benthic macrofauna data from the sediments of the Bay of Morlaix, sampled at 21 times over April 1977 to February 1982, covering the period of the Amoco-Cadiz oil tanker wreck in March 1978. This was last seen in Section 10 and introduced in Section 8 where the species-level *n*MDS (and *tm*MDS) showed the strong community change following the oil-spill and the subsequent partial recovery, with the re-establishment of a clear seasonal cycle. Open that workspace, Morlaix ws in C:\Examples v7\Morlaix macrofauna, or if unavailable, open the species data matrix Morlaix macrofauna abundance and the variable information (aggregation) file Morlaix macrofauna taxonomy. Calculate a couple of aggregation and transformation options, computing Bray-Curtis similarities and running *n*MDS, e.g. contrast plots for species-level, square-root transformed and family-level log transformed data (similarities sp-sqr and fm-log). The latter requires, on the active sheet Morlaix macrofauna abundance, **Tools>Aggregate>**(Variable information worksheet: Morlaix macrofauna taxonomy) & (From level: species) & (To level: family), followed by **Pre-treatment>Transform(overall)>**(Transformation: Log(X+1)) and resemblance etc as usual. On the resulting *n*MDS plot, take **Graph>Samp. Labels & Symbols** to remove labels and the (✓By factor) on symbols, and **Special>Overlays**>(✓Overlay trajectory: time). A similar pattern is seen to that for the species-level root-transformed case but showing an apparently greater degree of recovery. One possible explanation for this is seen in the line plots (*coherent curves*) of Section 10 – the effect of the highly abundant Ampelisca species prior to the spill, whose numbers crash and do not recover well, is more heavily down-weighted with the severe log transformation.

Second-stage *n*MDS (Morlaix macrofauna)

The illustration below has calculated all combinations of species (*sp*), genus (*gn*) and family (*fm*) level data, under no transform (*no*), square-root (*sqr*), fourth-root (*4th*), log(*x*+1) (*log*) transforms and reduction to presence/absence (*pa*), with similarity sheets sp-no to fm-pa. [Actually, all these have already been calculated for you, as PRIMER format *.sid files in the directory C:\Examples v7\Morlaix macrofauna\Morlaix similarities, and you can open as many of them as you need into the workspace in one batch by **File>Open**, highlighting them all and taking **Open**]. With one of these as the active matrix (it does not matter which – one of the very few routines for which that is true), run **Analyse>2STAGE>**(Data•Multiple matrices)>(Other resemblance matrices: ✓*fm-4th* & ✓*fm-log* & ✓*fm-no* & ...) & (Correlation method: Spearman rank). This returns a second-stage resemblance sheet of matrix correlations ρ, all of which are positive, with some very close to 1 (e.g. species and genus level under no transform; 4th root and log transforms for any of the taxon levels, etc), indicating robustness of the conclusions to those particular choices. Now run *n*MDS on this matrix to obtain the *2nd stage MDS* plot – note that the plot below has had its boundary shape changed with **Graph>Special>Main>**(Plot type•2D)>(Aspect ratio: 2.0). The main conclusions are that: transform choice and taxonomic level tend to have *orthogonal* effects (transformations run across the page, taxon levels run up the page); transform choice generally makes a larger difference to the outcome than taxon level (the exception being between 4th root and log, which are more or less equivalent – log being more severe than 4th root on very large abundances but less severe than 4th root for small counts); the difference between taxonomic levels increases with the severity of

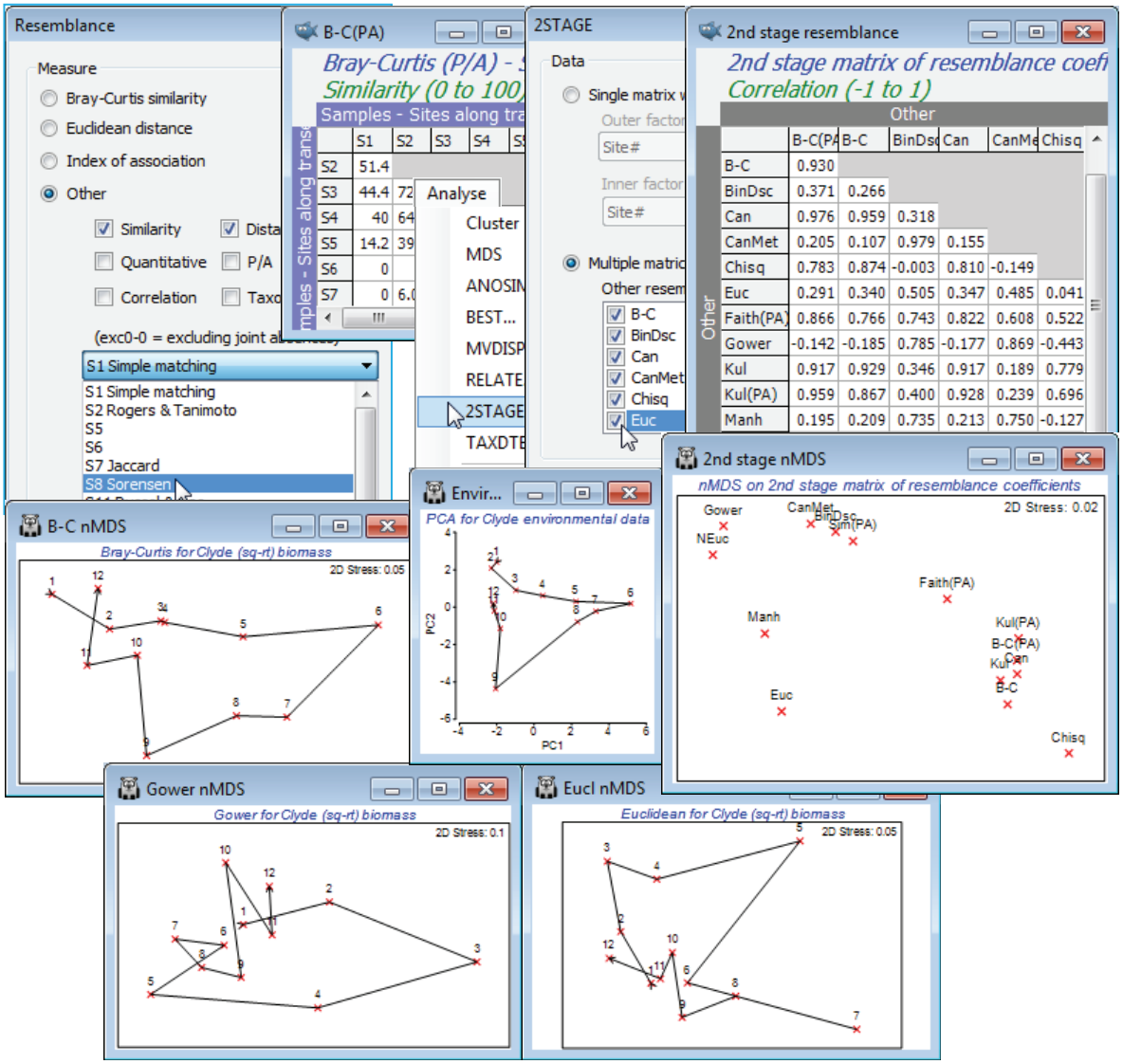
v7

the transformation. The latter is to be expected, since untransformed analysis tends to be dominated by a handful of species with the largest abundances – when these are in different genera or families their contribution is unchanged by aggregation. Save Morlaix ws for later this section, and close it.



2STAGE for resemblance coefficients (Clyde study)

The technique of 2nd stage plots has also been used (Clarke KR, Somerfield PJ, Chapman MG 2006, *J Exp Mar Biol Ecol* 330: 55-80) to examine the effects of different resemblance coefficient choices on a samples analysis, scaling this in relation to the effects of differing transformation (and, by extrapolation, taxonomic level). The environmental data from the Clyde sewage dump-ground study were used extensively in Sections 11 and 12 but, for this example, open the macrofaunal data C:\Examples v7\Clyde macrofauna\Clyde macrofauna biomass into a new workspace, and deselect the all-blank species (there are about 20 of them retained in this sheet because they have non-zero counts in the abundance matrix but their total biomass is too small to weigh). You can do this by **Select>Variables>**(•Use those that contribute at least 0.001%), then take a square-root transform (Biomass sq-rt). Always starting from this matrix, produce a wide range of distances and (dis) similarities using **Analyse>Resemblance>**(Analyse between•Samples) & (Measure•Other)> (✓Similarity) & (✓Distance/dissimilarity), selecting one coefficient at a time from the resulting list, e.g. *S1 Simple matching*, *S8 Sorensen* (i.e. Bray-Curtis P/A), *S13 Kulczynski (P/A)*, *S26 Faith (P/A), S15 Gower*, *S18 Kulczynski (quant)*, *Canberra similarity exc 0-0*, *D7 Manhattan distance*, *D10 Canberra metric*, *D16 Chi squared distance*, *Binomial Deviance (scaled)* and (•Bray-Curtis similarity) and (•Euclidean distance) from the main dialog, the latter both with and without normalisation of the species variables. [You should read the discussion in Section 5 and in Chapter 16 of CiMC on the suitability or otherwise of some of the coefficients in the full list, for non-count data]. With (say) *S8 Sorensen* as the active sheet, **Analyse>2STAGE>**(Data•Multiple matrices), ticking the check boxes for all the rest, and running *n*MDS on the resulting 2nd stage matrix. Look also at individual MDS plots for some measures with differing effect, in comparison with the contaminant gradient (the multivariate analysis for which – a PCA – is shown in Section 12). Crosses have been used for points in these plots, by changing the symbol type temporarily in the **Samp. Labels & Symbols>**(Default>Symbol:) dialog or, changing the global default by **Tools>Options>Graphs.**



Conclusions on comparing resemblance coefficients

Clarke KR, Somerfield PJ, Chapman MG 2006, *J Exp Mar Biol Ecol* 330: 55-80 discuss this analysis (and that for several other data sets) in more detail, but to pick out just four general points:
a) These 2nd stage plots have common features, irrespective of the actual data set, e.g. coefficients which are in what they term as the 'Bray-Curtis family' (including quantitative measures: S17, S18 & Ochiai (quant), matched by pres/abs measures: S8, S13, S14; also Canberra similarity exc 0-0) tend always to cluster on the 2nd stage plot, i.e. produce similar multivariate conclusions, and radically differ from Euclidean distance, even more so when the latter is normalised.
b) Choice of coefficient is much more crucial to a multivariate analysis than transformation (which itself is more important than taxonomic level – see earlier); this is apparent here by noting the relative proximity of the Bray-Curtis and Bray-Curtis P/A (Sorensen) points, and the Kulczynski and Kulczynski P/A points, on the 2nd stage plot (the first of the pair uses a mild square root, and the second is on presence/absence data – the most severe transform possible).
c) The inference of similarity from joint absences for coefficients such as Euclidean distance, S15 Gower etc, has a dramatically adverse effect on their performance in describing gradients of assemblage change where there is a turnover of species (i.e. pres/abs data is informative); this is clear from the above (1st stage) MDS based on Euclidean distance, which places site 6, at the centre of the dumpground, close to the extreme ends of the transect, 1 and 12, when 6 has no species in common with either! Similarity is deemed higher because they share absent species. The radical effect of counting (or not) joint absences is also clear here from: the separation of the Canberra metric from Canberra similarity (the only difference is an adjustment for double zeros, Section 5), and the way the plots splits left, right (counts 0-0, ignores 0-0), with the Faith coefficient intermediate since it counts joint absences, but with less weight than joint presences.
d) Another key feature which separates out the behaviour of coefficients is whether they implicitly or explicitly standardise (or normalise), and whether over samples or species. Chi-squared distance does both, removing all differences in total abundance between samples and also having a divisor of the total abundance of each species across all samples – low density species can be given very heavy weight, leading to problematic behaviour. Normalised Euclidean and Gower also have a species (but not sample) standardisation, giving rare and common species equal weight.

Close the workspace – we shall start a clear workspace next time we meet this data (Section 15).

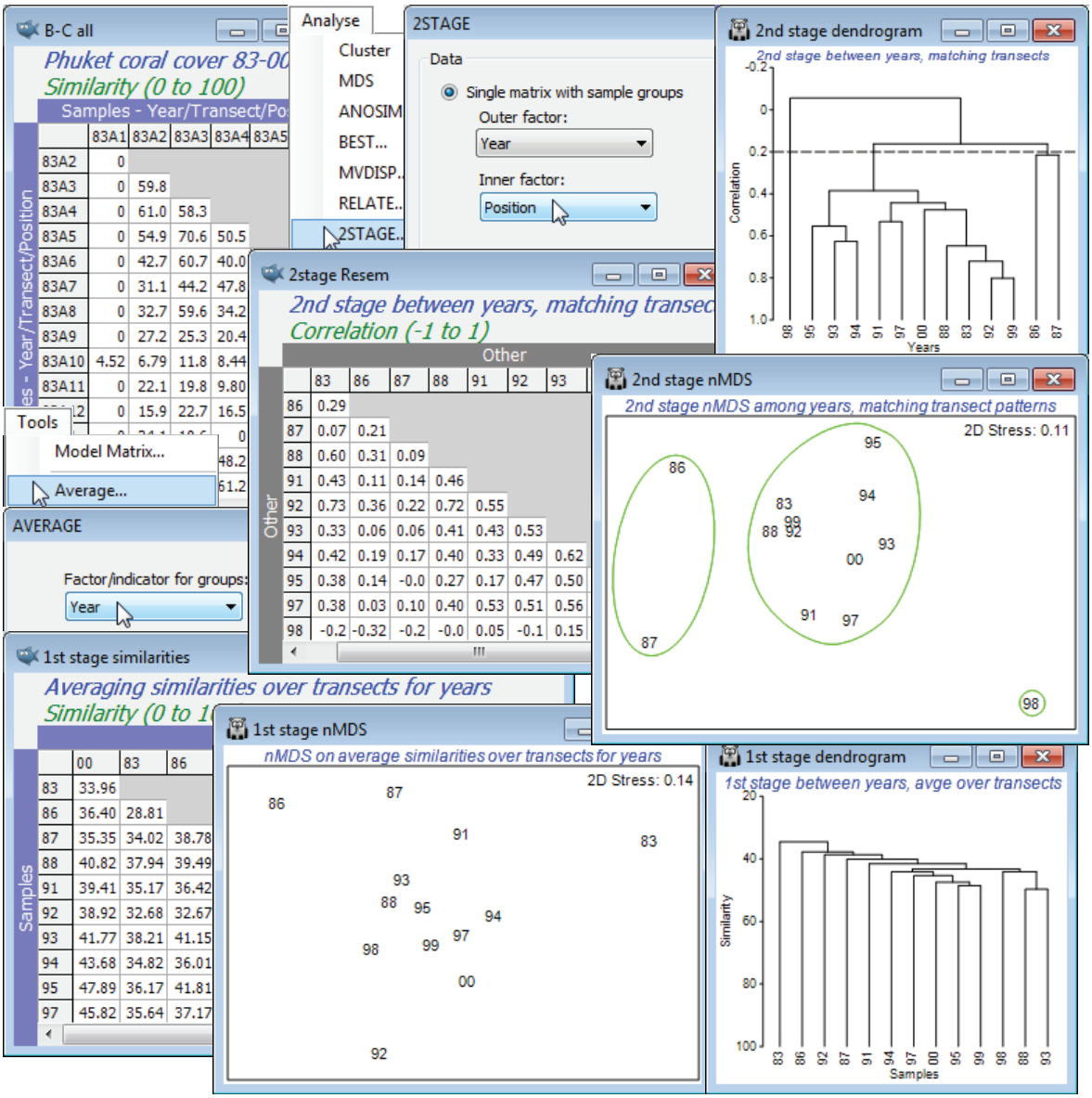
2STAGE for displaying 'interactions'

A very different way of using 2nd stage matrices is best accessed through the alternative entry option in the dialog box for **2STAGE**, namely to specify a single similarity matrix with factors defining a 2-way crossed layout of samples (e.g. of sites and times), and allow **2STAGE** to select the sub-matrices on which to calculate the second-stage correlations. To motivate this, return to the Phuket coral data at the start of this section, in which the spatial pattern of assemblage change over an onshore-offshore transect was compared for two years, 1983 and 1987. The rank correlation (Spearman) between the two Bray-Curtis similarity matrices underlying these profiles was only $\rho = 0.08$, indicating a poorly matching sequence, the conclusion being that the sedimentation from dredging for a deep-water port in 1986 and 87 had disrupted the spatial pattern of the assemblages. In fact, that study has data from 13 years over 1983 to 2000 (the merged file for which was created in Section 11). This period included a further potentially disruptive event in 1998, a prolonged high pressure anomaly creating a period of low sea levels, increasing the frequency of desiccation. If the transect patterns for *all* pairwise sets of years are now matched, a correlation matrix of ρ values is produced, which is the second stage matrix. These 'similarities' between years can be input to an MDS or clustering to give a visual summary of the inter-annual changes, not of the community as such (i.e. not of the average assemblage, or the assemblage at one fixed point on the transect – that would be a first-stage MDS) but of the internal pattern of assemblage change along the transect. Years which are anomalous in terms of their spatial pattern should stand out as outliers on this 2nd stage MDS or 2nd stage cluster analysis. If the inter-annual differences do not disrupt the internal spatial structuring but simply, for example, increase the abundance of all species down the transect in some years, relative to others, then the 2nd stage plot will show nothing whatsoever – that type of signal will be seen in a (1st stage) plot of yearly changes in the community, when averaged over the whole transect. In a sense, what the 2nd stage plot does is to remove 'main effects' of years (to use familiar univariate terminology) and concentrate on 'interactions', the changes in the internal spatial gradient for some years compared with others. This example is now implemented but is also discussed, along with other examples, in Clarke KR, Somerfield PJ, Airoldi L, Warwick RM 2006, *J Exp Mar Biol Ecol* 338: 179-192, and at the end of Chapter 16 of CiMC.

(Phuket coral transect)

Open the workspace Phuket ws, of coral cover for the Ko Phuket transect A, in C:\Examples v7\ Phuket corals, or if not available, open the data files Phuket coral cover 83-87, 88-97 and 98-00, and **Tools>Merge** them (as in Section 11), taking the defaults to produce the full inter-annual series, Phuket coral cover 83-00. This has 156 samples, in a 2-way crossed design split into 13 years, with 12 positions along the onshore-offshore transect (look at the factors Year and Position with **Edit>Factors**). Create the similarity matrix for all 156 samples as previously: Bray-Curtis on square-root transformed data, renaming it B-C all. On this, take **Analyse>2STAGE>**(Data•Single matrix with sample groups)>(Outer factor: Year) & (Inner factor: Position) & (Correlation method •Spearman rank) to produce the 2nd stage matrix, renamed 2stage Resem. On this, run **Analyse> CLUSTER** and **Analyse>MDS**, drawing clusters on the MDS using **Graph>Special>**(✓Overlay clusters) – see Section 8 – with slice at resemblance (ρ) of 0.2. Contrast this 2nd stage plot with the (1st stage) MDS of years, averaging over the transect positions with **Tools>Average** for factor Year (on original or transformed data, or perhaps from the similarity matrix B-C all – a case can be made for all three methods here!), then re-run the MDS. Although testing is impossible in this case, it is clear that this first stage plot of the year 'main effect' is less sensitive in picking up the impacts of sedimentation (86 and 87) and desiccation (98) than the second stage analysis, concentrating on the consistency over years of the spatial pattern along the transect ('interaction' effects).

v7



Save and close the Phuket ws. A more natural context for 2nd stage analysis is that of temporal studies, in which similarities in the time course are being compared across sites under different conditions, and this can give rise to cases where tests for this rather general concept of 'interaction' between time and space are possible, as in the time series for Tees Bay macrofauna now examined.

2STAGE for time series and repeated measures

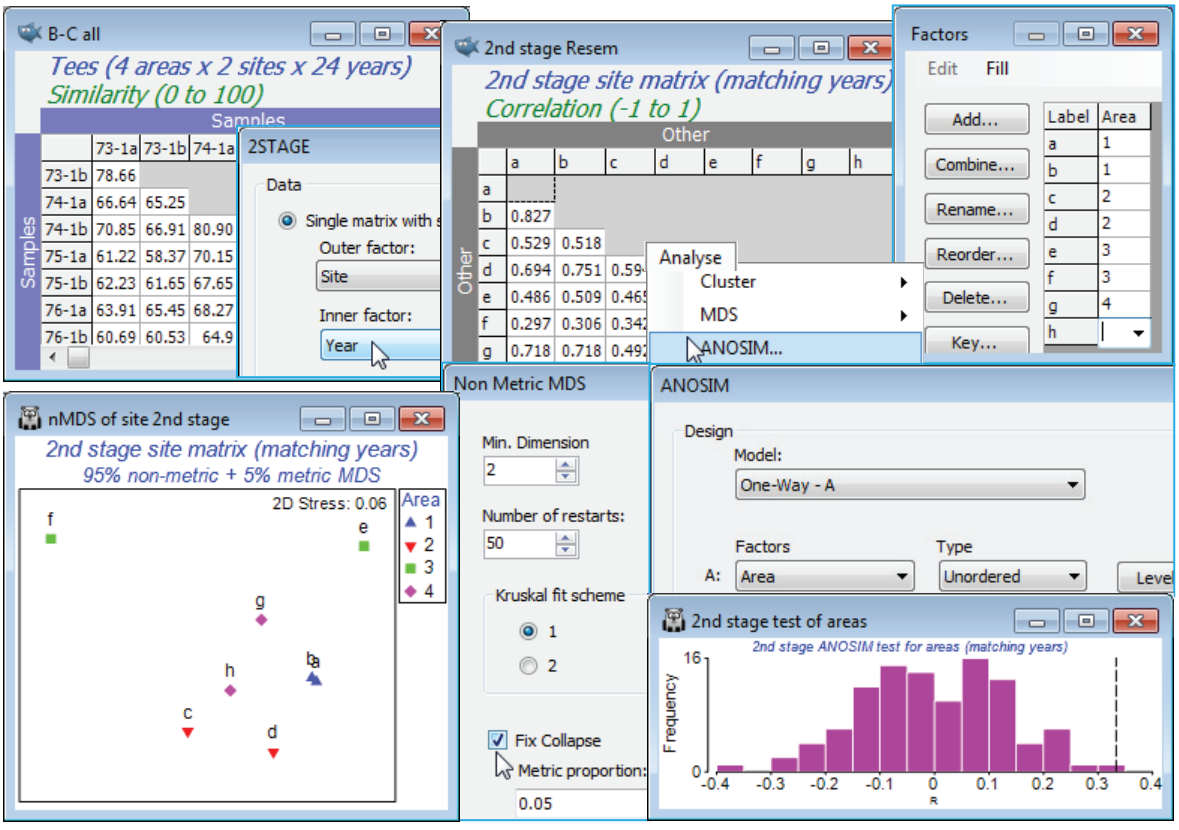
In the context of a 2-factor design, PRIMER makes a 2nd stage matrix very simple to produce but it is less easy to understand what it represents! The structure requires that the factors divide the data into a 2-way layout with no replicates in each cell; the inner factor specifies the patterns to match (spatial, for the Phuket data) and the outer factor is the one displayed (temporal, above). Note that, because of the symmetry of two-way crossed designs, these could be reversed, thus the Phuket data could have matched the inter-annual patterns at each point on the transect. This would remove the 'main effect' of differences in (time-averaged) assemblages along the transect, and concentrate on anomalous transect positions – those for which the relationship among years differs. The Clarke, Somerfield, Airoldi & Warwick 2006 paper, referred to above, discusses two further examples in which **Analyse>2STAGE** is able to match temporal patterns to produce a spatial second-stage matrix. Both have a natural hypothesis testing framework, which extends to *repeated measures* designs, usually considered problematic even in univariate studies. An inter-annual time series (1973-96) for subtidal macrobenthos, at two sites in each of four different areas in Tees Bay, UK, was met in Section 9, and will be exemplified here, and a repeated measures recolonisation study on macroalgae at Calafuria in the Ligurian Sea (the non-repeated measures data from which was seen in the ANOSIM section) is also discussed in detail as the last example in Chapter 16, CiMC.

(Tees Bay macrofauna)

The workspace Tees ws was saved in Section 9; if not available open the data Tees macrobenthic abundance from C:\Examples v7\Tees macrobenthos and recalculate Bray-Curtis similarity on the 4th-root transformed abundances for all 192 samples (B-C all), with structure of 4 areas (1-4) in each of which the same 2 sites (a,b; c,d; e,f; g,h) were sampled in September over 24 years (1973-96), with each sample a pool of a consistent number of benthic grabs. A question of interest here is whether the areas show the same inter-annual patterns – as might be expected if they are primarily influenced by wide-scale climatic variation – or whether local factors, such as the proximity of the plume from the Tees estuary to some areas (with the inevitable local changes in an industrialised estuary) result in different time trajectories in different areas. This can be addressed by calculating, on B-C all, a second-stage matrix for the 8 sites, by **Analyse>2STAGE>**(Data•Single matrix with sample groups)>(Outer factor: Site) & (Inner factor: Year). On the resulting sheet 2nd stage Resem, set up the simple Area factor (by **Edit>Factor>Add**) with entries 1,1,2,2,3,3,4,4 – the sites are the replicates – and run 1-way ANOSIM on factor Area, and *n*MDS plots of the 8 points (you will need the ✓Fix collapse option, Section 8). The ANOSIM is not a test for different assemblages over the areas – that is inevitable given the spatial range – but removes those, and shows that the temporal variations for each area (from different baseline communities) are not the same (R = 0.33, p<1%).

v7

v7



(Calafuria macroalgae experiment)

The Calafuria macroalgal recolonisation experiment monitored the same physical rock patches over one year, having first cleared the (subtidal) rockface. Replicate patches were tracked for 8 different 'treatments', namely different times of year for the clearance. The 2STAGE analysis matches the recolonisation time patterns of all replicates, and a 1-way ANOSIM on the 2nd stage matrix tests whether different treatments give different recolonisation profiles (which they do). The individual time points in the recovery sequence cannot be assumed independent, since the same rock patch is returned to bi-monthly – this is *repeated measures*. But the 2nd stage analysis treats that inter-dependent time sequence of recovery as a single experimental unit, in effect. It becomes a single point on the 2nd stage MDS plot and a single replicate in the 2nd stage ANOSIM, independent of other replicates (other rock patches), and thus gives a fully valid test. An equally valid alternative would have been to throw away the intermediate recovery times and just analyse the assemblages at one year after clearance (which is the data analysed in Section 9, which also introduces a lower level to the design, of plots within areas, under the different treatments). In fact, the second-stage analysis is more incisive here because it allows the whole recovery profile to be assessed rather than solely its end point – but different hypotheses are being tested, and both are of interest.

Other BEST applications

Another situation employing rank correlation (ρ) between two resemblance matrices is the **BEST** (*Bio-Env*) routine of Section 13, where the biological similarity matrix ('response') describes the among-sample relationships of the full community and the secondary data sheet ('explanation') is of environmental variables. Subsets of the latter variables were taken, and among-sample distances computed for each subset and correlated with the biotic similarities, the search being for a variable set that maximises ρ. However, there is nothing in the construction of BEST which limits its use to species similarities and environmental matrices. Either or both of these two sheets could be from biotic or abiotic samples – the user needs only to specify a resemblance measure which is relevant for the type of data in the secondary data matrix. A number of possibilities can be envisaged. In what might be termed *Env-Bio*, subsets of species could be selected which best characterise the environmental gradient defined by a specified set of abiotic variables, or best match a simple model structure, e.g. the seriation distance matrix for *n* equally-spaced points on a line, as in the Phuket corals transect example earlier in this section ("which species define the serial gradient along the transect?"). Or for samples which have an *a priori* (unordered) group structure, a relevant model matrix of distances was seen to consist simply of 0's (within groups) and 1's (among groups). An *Env-Bio* analysis in that case would search for subsets of species which, in combination, best split the samples into those pre-defined groups – a rather different form of SIMPER analysis (Section 10) acting on all the groups at once, rather than selected pairs. It is equivalent to optimising the ANOSIM R statistic, PRIMER's preferred measure of group separation in high-d space. [We saw ANOSIM R used in the same way earlier, Sections 6 & 13, in searching for optimal subdivisions of samples in divisive clustering, though there the set of species was fixed and the sample divisions selected, and here the sample groups are fixed and the set of species is being searched over. It should be stressed again that having selected an optimal species set, it is totally invalid to re-test the groups with a simple ANOSIM test! The strong selection bias effect is allowed for, however, in the global BEST test of Section 13, so that when sample groups are fixed *a priori* the BEST test could be used to justify interpreting the selected optimal species subset as 'better than chance'.]

A further generalisation would allow ordering on the groups, e.g. for the *seriation with replication* model matrix described earlier in this section. There the idea would be to select the subset of species which best characterise an ordered group structure of community change, i.e. lead to both good separation of the groups from each other and in their pre-defined order (e.g. as in the distance groups for the Ekofisk oil-field study). A similar use of variable selection to best match *a priori* ordered groups was given by Valesini F *et al* 2003. *Est Coast Shelf Sci* 57: 163-177, under what might be termed an *Env-Env* scenario, since the variables were beach morphology characteristics, and thus required a distance-based resemblance calculation, such as normalised Euclidean. Other natural applications of this type might include the selection of biomarkers to best display a given impact gradient determined by tissue chemistry, the selection of morphometric measurements to best characterise known species or sub-species categories (unordered groups or ordered clines) etc, again supplemented by the global BEST test, to allow for the selection bias when testing overall significance of the 'explanation' (but see the important reservations expressed in Chapters 11 and 12 of CiMC on the extent to which correlative-type links of species to environmental variables, biomarkers to tissue contaminants etc, are ever demonstrated to be causal).

BVStep stepwise selection

There is one fundamental problem with applying BEST (Bio-Env) in many of the above scenarios: the number of variable *combinations* from the active matrix that must be considered in a full search increases exponentially with the number of variables. For p variables, there are $(2^p - 1)$ combinations, and this is prohibitive for p more than about 16 (c. 65,000 combinations). Searching across all subsets of species from a typical community matrix will therefore usually prove impossible. The (•BVSTEP) option under **Analyse>BEST** instead carries out a stepwise search: the best single variable is selected (maximising the matching coefficient, ρ); this is retained and the best variable to add to this is selected (maximising ρ); these two are retained and a third variable is added, and so on, resulting in a declining number of combinations to be considered at each step. This is called *forward selection*. BVStep also carries out *backward elimination*: starting with all the variables included, the one that decreases ρ least, when omitted, is dropped from the set, and this elimination process repeated. In fact, as is common with stepwise procedures elsewhere (e.g. in multiple linear regression), BVStep implements both forward and backward steps successively, so that after each addition of a variable by forward selection, the current set of variables is scanned to see if any of the other variables can now be eliminated. (The analogy with stepwise multiple regression is not perfect, note, because there the residual sums of squares always decreases as more variables are added – here the ρ value may go up or *down*, giving a natural optimisation). It follows, however, from the fact that only a small fraction of the possible combinations are considered, that the routine can become trapped in a non-optimal maximum, just as *n*MDS can get trapped in a local minimum of the stress function (Section 8). The answer is the same as for MDS – repeat the search from a different starting position. So, the BVSTEP dialog lets the user specify how many random restarts are required (choose as many as are computationally feasible). Each restart is from a different, randomly chosen, combination of the variables – experience suggests that it is better not to start with too large a number because it can be difficult to shed extremely sparse variables that neither help nor harm the best solution, so the default is set at 6. Chapter 16 of CiMC gives more detail on the operation of the forward/backward stepping algorithm and the application below.

Species sets 'explaining' the overall pattern

The main application area for the BVStep routine introduced by Clarke KR & Warwick RM 1998, *Oecologia* 113: 278-289, is what might be termed *Bio-Bio*, namely searching for subsets of species whose resemblance matrix best matches that of another (fixed) set of species. One can envisage this used on different faunal (taxonomic- or trophic-based) groups to elucidate potential interactions but the most obvious context is when the two biological matrices are from the same data. That is, the input similarity matrix is computed from the full set of species, and the secondary data sheet from which species are selected is the same full species data. Now, the idea is not to maximise ρ, since it can always be made equal to 1 by choosing a subset which is the full set of species, but to find the smallest possible subset of species which, in combination, describe most of the pattern in the full data set. 'Most' in this context is taken to be a conventional, and somewhat arbitrary, $\rho > 0.95$. Once ρ gets to about this level, two multivariate patterns (e.g. as seen in 2-d ordinations) are effectively indistinguishable, and would not lead to different interpretations.

The procedure can be thought of as a generalisation of the SIMPER approach (Section 10) to the case of continuous multivariate patterns, rather than a clearly-defined clustering of samples. For example, in the Morlaix MDS of the time series of 21 samples, seen earlier in this section, SIMPER could perhaps be run on three groups of times – before and immediately after the oil-spill, and the partial recovery phase, to identify all species contributing to the dissimilarity between each pair of those groups. The BVStep procedure, however, asks a subtly different question, namely, is there a subset of species which between them account for the whole continuous pattern: the structure of initial seasonal cycle, a period of marked change following the oil-spill, then a gradual recovery with the re-establishment of the seasonal cycle? Not only does this provide a more holistic answer than SIMPER (and, importantly, one that can be applied whatever the chosen resemblance matrix), it is also more parsimonious in identifying indicator species: if several species are contributing to the pattern in exactly the same way, BVStep will only need to select one of them, but SIMPER will identify all as contributing something to the average between-group dissimilarity. A next question is then to ask whether the identified set of species is the only subset which is capable of accounting for this multivariate impact, recovery and seasonal pattern (i.e. would constitute a good set of indicators for this time series). In other words, is the same pattern reinforced in the matrix over several sets of species? – what might be termed *structural redundancy.*

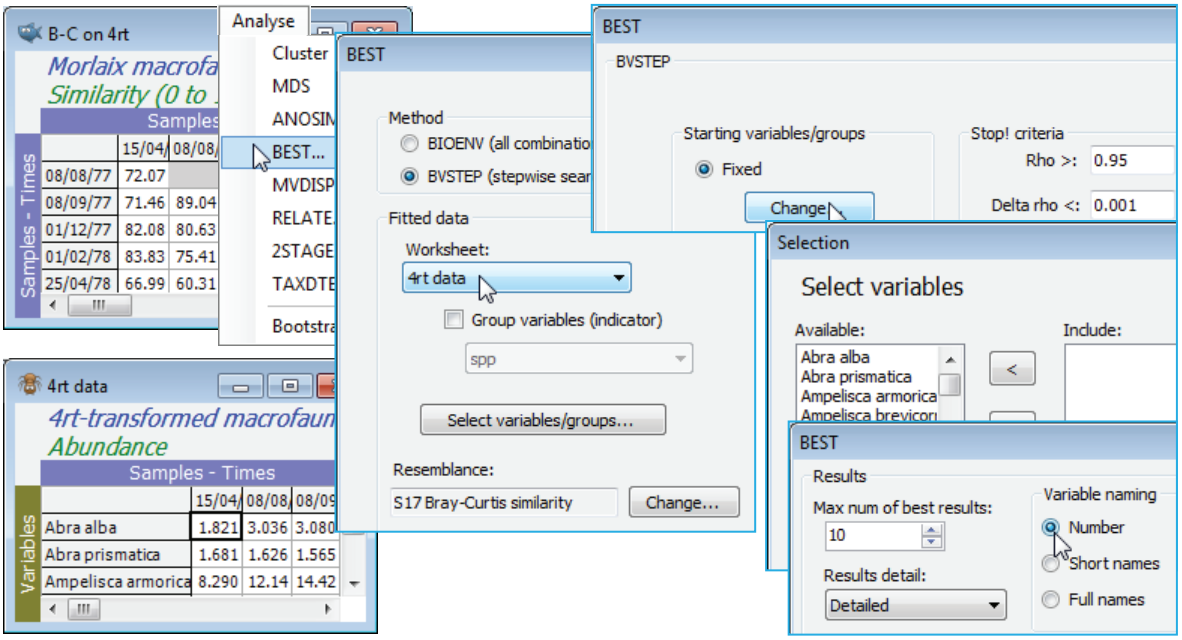
BVStep (Morlaix macrofauna)

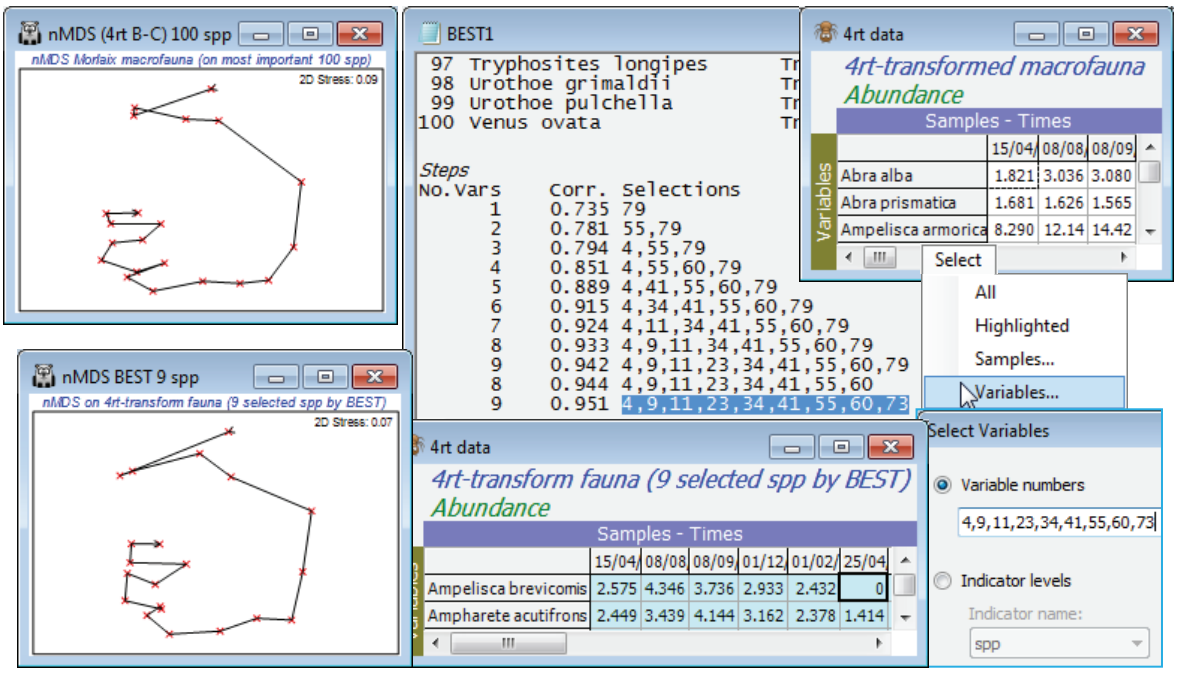
Re-open the Morlaix ws workspace in C:\Examples v7\Morlaix macrofauna from earlier in this section, or since this is all that is needed, just open the data file Morlaix macrofauna abundance into a clear workspace. It consists of 21 sampling times and 251species. Clarke & Warwick 1998 reasoned that many of these species were sufficiently rare (over half have totals across all samples in single figures) that the problem could be scaled down by removing those – so reduce to the *most important* 100 (see Section 3). Thus, **Select>Variables>**(•Use n-most important where n is 100) on Morlaix macrofauna abundance, then fourth-root transform, naming it 4rt data. (A severe transform seems the best choice, otherwise the counts of tens of thousands in a few species will dominate, as can be seen from a shade plot). Generate the *n*MDS ordination from Bray-Curtis similarities on this reduced, transformed data, calling the resemblance matrix B-C on 4rt. This is the active sheet on entry to **Analyse>BEST**, which takes the transformed data matrix 4rt data as its secondary sheet and searches for the smallest possible subset of the 125 species that effectively contains (to within ρ>0.95) the same among-sample information as B-C on 4rt. It is clear that the full enumeration of possibilities in the (•BIOENV) option would never be possible (2^{100} species combinations!) so the stepwise option of (•BVSTEP) is necessary. Even with the reduction of species numbers, it must be realised that many of these 100 species will be highly inter-correlated, and it is inevitable that many marginally different combinations of species will do an almost equally good job as indicators of the full data set (a point also made in Section 13 about linking biotic and abiotic variables). It is desirable therefore to start the search from several random subsets (perhaps 50), and look at all the output results (Detailed) – if only to appreciate that we are very far from having a single 'correct' answer! Nonetheless, it is interesting to see that the detailed MDS based on 100 species can be reproduced almost perfectly by several competing selections of only 8 or 9 species, as follows.

BVStep starting and stopping options

On B-C on 4rt, **Analyse>BEST>**(Method•BVSTEP) & (Worksheet: 4rt data), taking the defaults for all other entries (Spearman correlations, the suggested Bray-Curtis similarity, all 100 species Available for selection, and the permutation test ignored – a test of $\rho = 0$ makes no sense in this context and is invalid when the same data are being used in both matrices). On the **Next>** dialog, for BVSTEP options, take (Starting variables/groups•Fixed), i.e. on the **Change** button, no species are in the Include category, so the stepwise routine starts from no species and forward steps. An alternative is to Include them all and the routine will then work largely in backward elimination mode, though – as previously mentioned – this tends not to work as well since it can be difficult to drop species that are so sparse that they add or detract nothing. The (Stop! Criteria Rho>: 0.95 & Delta rho<: 0.001) choice ensures that the routine will keep searching until either the improvement in ρ at the next step is <0.001 or the cut-off for acceptable ρ of 0.95 is reached. On the final dialog, take (Results detail: Detailed) & (Variable naming•Number). Use of •Short or •Full names makes it easier to immediately identify the species, but numbers have the advantage that a species list can be copied/pasted from the Results window to the **Select>Variables>**(•Variable numbers) box, so that the optimal species set can easily be extracted from 4rt data and the similarity and MDS re-run.

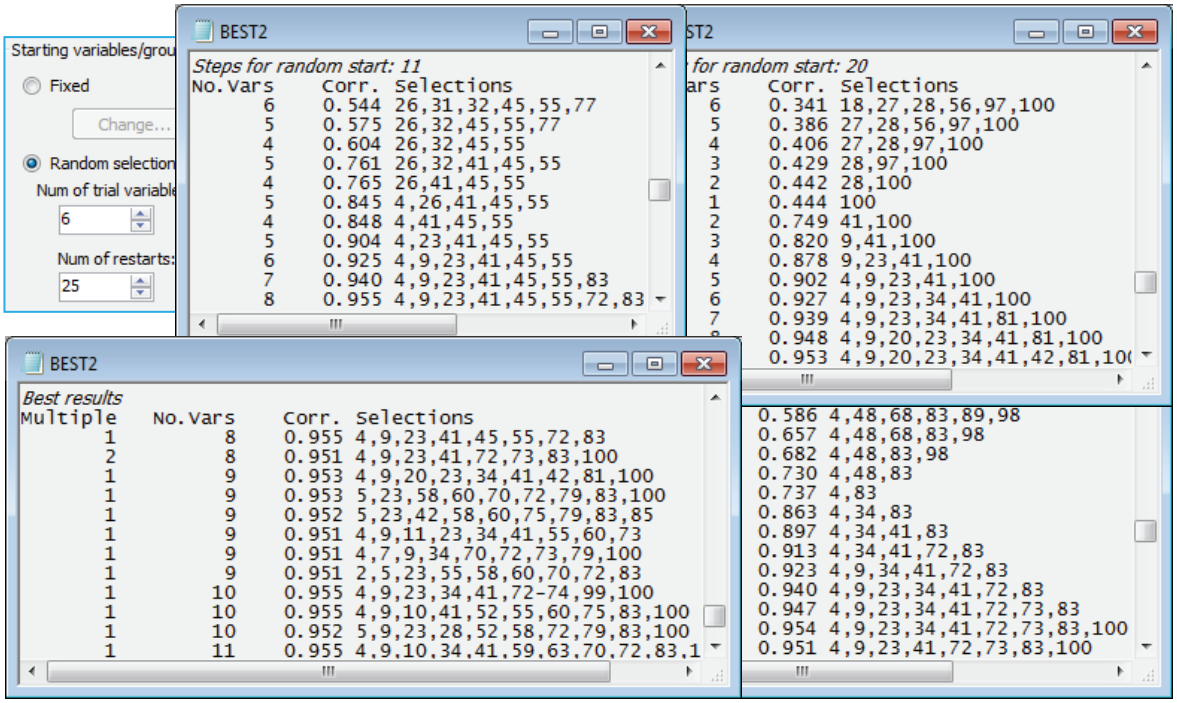
v7





BVStep from random starts

Starting the iterative search process from a blank species list is certainly not guaranteed to get you to the best solution (minimum number of species which give ρ≥0.95) – it is easy to get trapped in a local optimum which in not the globally best solution (which can never be known for certain). In fact a marginally better solution, in the sense of involving only 8 species variables, can be found in this case. See this by re-running the routine from different starting places, having first reinstated the full 100-species transformed data matrix 4rt data by **Select>All** – this is important otherwise you will find yourself searching only through the 9 species! (You may also wish to remove highlights with **Edit>Clear Highlight**, though this is not important). The first dialog for **Analyse>BEST** (run on active sheet B-C on 4rt) is the same as previously, but on the BVSTEP dialog take (Starting variables/groups•Random selection)>(Num of trial variables/groups: 6) & (Num of restarts: 25). This starts the stepwise routine from a randomly chosen 6 species from the 100, and (Results detail: Detailed) and (Max num of best results: 25) on the last dialog will allow you to see the alternating backward elimination and forward stepping phases in the Results window. It also permits the final (*Best Results*) table to display all the solutions obtained – in best to worst order – in the event that they are all different (which they nearly all are in the case below!).



Remember that these are not listed primarily in decreasing order of ρ but in increasing order of the number of species. Only when two sets have the same number of species is the ρ value (which has to be ≥0.95 for that solution to be listed at all) taken into account. You will obtain a different set than this (though probably overlapping), since a differing random number seed is used to select the starting species in every new run of the program. Occasionally, the search will end prematurely before 0.95 is reached, even though we know a value of 0.95 exists if we are searching the whole matrix (a value of 1 then exists!) – in that case try using Delta rho<: 0.0001, or even smaller, to try to keep the addition and deletion of species operating, and/or increase the number of restarts. The second-best solution above was found twice (see the *Multiple* column in the *Best results* table) but many more than 25 restarts would probably be needed to be reasonably content that a 7-species solution could not be found. Setting out on an exhaustive search here rather misses the main point, though, that the impact and seasonal structure in the above MDS – which, importantly, is largely 'signal' because of the large sample sizes (we are not chasing 'noise') – can be displayed in just the same way by a small set of 8 or 9 species. A close look at the near-optimum solutions shows that many of the same species are involved in several of these.

In the further analysis in Fig. 16.3 of CiMC, from the Clarke and Warwick 1998 *Oecologia* 113 paper (and based on a somewhat larger number of species retained from the original *c*. 250), **BEST** is re-run, excluding this first subset of BEST-selected species, but again matching to the B-C on 4rt similarities from the retained set. In the above case, we therefore need to exclude species 4, 9, 23, 41, 45, 55, 72, 83 from the 4rt data matrix before entering it as the (secondary) worksheet to BEST. The quickest way of doing this, as seen earlier, is to copy and paste those 8 numbers to the **Select> Variables>**(•Variable numbers) box, then **Select>All** to leave them highlighted, and **Edit>Invert Highlighted** followed by **Select>Highlighted** will leave the remainder of the transformed 4rt data matrix selected for this second run of **BEST**. The matrix then needs to be **Tools>Duplicate**(d) in order to use the same trick to remove both the first and second species sets selected by BEST, if a third species set is sought. For the Morlaix data, about five entirely separate *species peels* can be found, all of which essentially reproduce the same multivariate pattern, indicating a high level of structural redundancy in the matrix.

v7

This is the 'opposite side of the coin' from the *coherent curves* analysis we saw for this data in Section 10, where species were grouped into (about 8 or 9) distinct and characteristic sets in terms of their temporal patterns, seasonally and in response to the oil-spill and its aftermath. Each set contains several species which are able to substitute for each other, in the sense that their time patterns are statistically indistinguishable. Conceptually, it should be the combination of the (numerically more dominant) species drawn from each of these sets which tend to make up the above species peels, between them representing the range of temporal responses and therefore capable of recreating the community pattern for the full data set.

Close the Morlaix workspace – it will not be needed again.

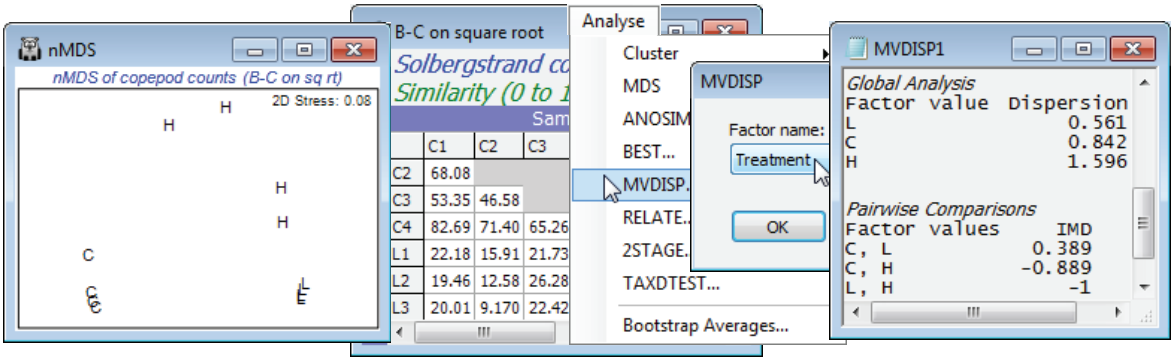
Multivariate dispersion MVDISP

One of the few multivariate routines not so far met is **Analyse>MVDISP**, applied to a resemblance matrix from samples with a simple group structure (i.e. a 1-way layout, or a crossed design that can be turned into a 1-way layout by defining an appropriate combined factor). This gives a description of relative multivariate variability within each of the groups in a single ordination or, to be more precise, in the full-dimensional space of the rank similarity matrix underlying that ordination. (As such it is not a matching of multivariate patterns and doesn't really belong in this section – but it has to go somewhere!) The concept is again a simple non-parametric one – though rather limited in scope – and described in the Increased Variability section of Chapter 15 in CiMC, so only an example will be given here. Tables of the *dispersion sequence* of all groups (equation 15.4) and the *index of multivariate dispersion (IMD)*, comparing pairs of groups (equation 15.2), are output to the results window, and these measure differing relative dispersion across groups on the basis of dissimilarity (or any other resemblance measure) within groups – between-group dissimilarities are not used. [If a description retaining the actual similarity scale is required, an alternative – for the special cases of Bray-Curtis and Euclidean distance – might be to run **Analyse>SIMPER** on the transformed data sheet and look at the headings for the first set of tables, each of which gives the average similarity of all pairs of replicates within that group. More generally – for all other coeff-icients – the same information about average resemblance within groups is one of the tables output from the PERMANOVA+ add-on routine, PERMDISP, see the Anderson *et al* 2008 user manual].

The term *multivariate dispersion* rather than *variance* is used because the relationship between the univariate variance of the original variables and the dispersion in 'resemblance space' (and its low-dimensional ordinations) can be far from linear, depending on the choice of resemblance measure. For example, similarity measures in the quantitative Bray-Curtis family (see earlier this section on comparison of resemblance coefficients) are driven partly by the presence/absence structure of the data, as well as the magnitude of counts from species which are always present, and this inevitably involves a non-linear transformation of original variable scales. Similarly, something as simple as normalisation, used in a Euclidean distance analysis of environmental variables, will remove any direct link between variance on the original measurement scales and dispersion in the multivariate space. Any statement about relative dispersion, therefore, must be contingent on specifying the resemblance measure used. Clarke KR, Somerfield PJ, Chapman MG 2006 *J Exp Mar Biol Ecol* 330: 55-80, show the radically different conclusions that would be reached, for the Tikus Island reef coral study (met in Section 5), about the dispersion among transects before and after a coral bleaching event, under Chi-squared, Bray-Curtis (standard and zero-adjusted) and Euclidean-based analyses – with the intermediate (zero-adjusted) Bray-Curtis arguably giving the most informative description, in terms of identifying the inter-annual changes in coral community. (This interesting example is also extensively discussed in Chapter 16, CiMC, on resemblance coefficient choice.)

(Mesocosm experiment, Solbergstrand copepods)

The illustration used here is a simple 1-way design of 3 mesocosm treatments: Control (C), Low (L) and High (H) dose of organic enrichment applied to the surface of 12 intact sediment cores, taken from the same location into a mesocosm system, and randomly allocated to the treatments (with 4 replicates in each). Data are from Gee JM *et al* 1985 *J Exp Mar Biol Ecol* 91: 247-262, as analysed in a multivariate way by Warwick RM & Clarke KR 1993 "Increased variability as a symptom of stress in marine communities" *J Exp Mar Biol Ecol* 172: 215-226. Chapter 15, CiMC shows analysis of the resulting meiofaunal communities in the sediment cores (nematodes and copepods) after several weeks' exposure, but here we open just the copepod data, Solbergstrand copepod counts in C:\Examples v7\Solberg copepods. For square-root transformed data and Bray-Curtis similarities, plot the *n*MDS and note the apparently much larger dispersion within the High dose treatment (as well as the obvious differences between treatments, which would be tested, validly, by 1-way **ANOSIM**). This is indicated more reliably, i.e. not in the low-d approximation of an ordination plot, by running **Analyse>MVDISP>**(Factor name: Treatment) on the resemblance matrix. The dispersion sequence of 0.56, 0.84, 1.60 for L, C, H shows that the average rank dissimilarity is almost three times higher within H than L (comparable dispersions result in a sequence of 1's), and the pairwise comparisons show that all the lowest dissimilarities (within a group) are in L and all the highest in H (thus *IMD* = –1 for that pair of treatments). The result, however, is of limited usefulness since an exact permutation test of these dispersion differences is not possible under the non-parametric framework in PRIMER, for much the same reason as interaction tests in a two-way crossed layout are not possible, see the comments at the end of Section 9 and Chapter 6 of CiMC. [No permutation procedure exists under a null hypothesis that the dispersions are the same for each group, but that the 'locations' – in so far as they are defined for rank-based dissimilarities – may differ. If the primary interest is in testing for differences in multivariate dispersion of groups, for a given resemblance measure, you should use the (approximate, semi-parametric) permutation test given by the PERMDISP routine in PERMANOVA+ – see the Anderson *et al* 2008 manual. The parameters defining centroids of each group in the high-d PCO space are estimated and each centroid is moved to the same point, justifying permutation of the samples across groups under the null hypothesis – if location differences have been removed, and the null hypothesis specifies no dispersion differences, then sample labels again become interchangeable.]



15. Biodiversity measures and tests (*DIVERSE, TAXDTEST*)

Input/output for diversity

PRIMER computes an extensive set of univariate diversity measures, covering most of the standard indices used in ecology. The active sheet is a data matrix for which the chosen indices are calculated for every sample. The measures are selected by ticking check boxes, so any combination of them can be computed in one run, and the results output either to the results window in a tabular format (which can be copied to the clipboard and pasted directly into Excel) or as a samples-by-variables matrix in a second worksheet. The latter can be saved, as usual, in text or Excel format, for transfer to a standard univariate stats package, but PRIMER 7 can now produce means and confidence interval plots for sets of univariate data, and the PERMANOVA+ add-on can perform permutation-based ANOVA on each variable (univariate being a special case of multivariate).

v7

Presentation of diversity information

The facility to send the indices to a new worksheet also allows some interesting possibilities for further presentation, including multivariate analysis. For example, the indices can be superimposed, one at a time, on an MDS plot for the full species assemblage data (treat the diversity matrix like an environmental variables data file) or input the diversity matrix to a multivariate analysis itself (again treat the indices as an environmental array and calculate normalised Euclidean distances between samples for an MDS, or run a PCA). This will show the between-sample relationships obtained from the full range of diversity information extracted, and can be contrasted with the usual ordination exploiting the matching of species identities between samples (which is generally found to be more sensitive, since it exploits more of the available information). A PCA for a large set of diversity indices can also demonstrate how many genuinely different axes of information they have captured (i.e. how many PC axes explain most of the variability), since many standard indices are really just some weighted combination of two features: the total number of species (richness) and the extent to which the total abundance is spread equally amongst the observed species (evenness). An MDS plot of the variables, using (absolute) correlations between indices as the resemblances (an analysis mentioned previously for species, but considered likely to be too high a stress there to be useful) is now viable and shows which measures are essentially equivalent. Such analyses can be an incentive not to proliferate indices by defining yet further variations of the same information.

Taxonomic distinctness

One of the distinctive features of PRIMER is its inclusion of a suite of biodiversity measures based on the relatedness of species within a sample, e.g. the average 'distance apart' of any two species or individuals chosen at random from the sample (termed *average taxonomic distinctness*). This is usually defined from a Linnaean tree (though could be phylogenetic, genetic or functionally-based) and requires availability either of an aggregation file (Section 11) covering all the species in the data matrix, which will be used to compute species distances, or a more direct species resemblance matrix, supplying genetic or functional distances among species. It provides an added dimension of information to that obtainable from the abundance distribution alone: as an average measure its construction makes it independent of the number of species, and it thus has much better statistical sampling properties than richness-related estimators when sampling effort is non-comparable over samples. This should be seen as the major sphere of application: uncontrolled studies over wide spatial or temporal scales, where classic diversity measures can be misleading. Several papers describe the methods, e.g. Clarke KR & Warwick RM 1998, *J Appl Ecol* 35: 523-531, Clarke KR & Warwick RM 2001, *Mar Ecol Prog Ser* 216: 265-278 and Warwick RM & Clarke KR 2001, *Oceanog Mar Biol Ann Rev* 39: 207-231. A detailed exposition is also given in Chapter 17, CiMC.

v7

In just the same way as for the classic indices, PRIMER can calculate a range of such taxonomic-related measures (including the *PD* of Faith DP 1992, *Biol Conserv* 61: 1-10), through check boxes on the **Analyse>DIVERSE** menu. These can be separated into quantitative indices (e.g. Δ, Δ^*) and those which depend only on a species list (indicated by a superscript +). The latter are divided into average measures (e.g. Δ^+, Λ^+) which have the property of independence of sampling effort (in their mean values), and total measures (e.g. $S\Delta^+$, $S\Phi^+$) which are alternative definitions of the taxonomic richness, combining the number of species with relatedness information. For two of the presence/absence measures, a hypothesis testing structure can be erected to compare a location's observed *average taxonomic distinctness* (AvTD, Δ^+) and *variation in taxonomic distinctness* (VarTD, Λ^+) with that 'expected' from a regional master list, assuming assembly rules for the species set which are independent of their taxonomic inter-relation. This is run by **Analyse>TAXDTEST**, when the active window is either an aggregation file or a variable (dis)similarity matrix.

v7

Standard indices calculated

The range of indices available is illustrated with the macrobenthic data Clyde macrofauna counts from the Clyde sludge dump-ground study, directory C:\Examples v7\Clyde macrofauna, last seen in Section 14. Analyses so far have used only the abiotic and biomass matrices, and the existing workspace Clyde ws may have become cluttered, so open Clyde macrofauna counts into a new workspace, and save it as Clyde ws2. Without pre-treatment, take **Analyse>DIVERSE>**(✓Results to worksheet). Look at the options on the first 5 tabs, taking only ✓S, ✓d, ✓J′, ✓H, ✓α, ✓H′ (log base e), ✓1 – λ′, ✓ES(n) with n values: 15, 30, 45 (there is no special significance to the index grouping under tabs, except that the last two tabs deal with taxonomic-relatedness measures, seen later). The abundance of the *i*th species is denoted by N_i (i = 1, 2, .., S) and, as a ratio of their sum (N), this is denoted P_i (i = 1, 2, .., S). The first 5 tabs (where ✓ denotes the default selections) are:

Other

✓Total species: S
✓Total individuals: N
✓Species richness (Margalef): $d = (S - 1)/\log_e N$
✓Pielou's evenness: $J' = H'/\log_e S$
Brillouin: $H = N^{-1} \log_e\{N!/(N_1!N_2!\ldots N_S!)\}$
Fisher's α statistic

Shannon

✓$H' = -\Sigma\, P_i \log(P_i)$, where the logs are to the base e
H' as above but for logs to the base 2
H' as above but for logs to the base 10

Simpson

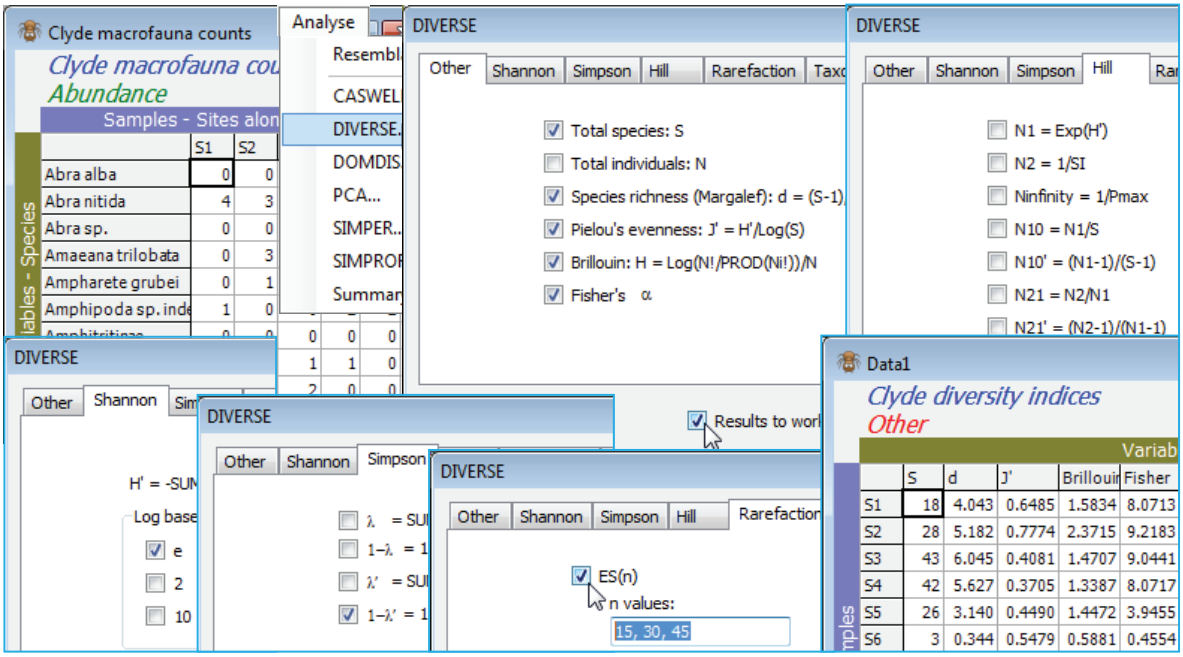
$\lambda = \Sigma\, P_i^2$
$1-\lambda = 1 - (\Sigma\, P_i^2)$
$\lambda' = \{\Sigma_i\, N_i(N_i-1)\}/\{N(N-1)\}$
✓$1-\lambda' = 1 - \{\Sigma_i\, N_i(N_i-1)\}/\{N(N-1)\}$

Hill numbers

$N1 = \exp(H')$
$N2 = 1/\Sigma\, P_i^2$
$N_\infty = 1/\max_i\{P_i\}$
$N_{10} = N1/S$
N_{10}' = $(N1-1)/(S-1)$
$N_{21} = N2/N1$
N_{21}' = $(N2-1)/(N1-1)$

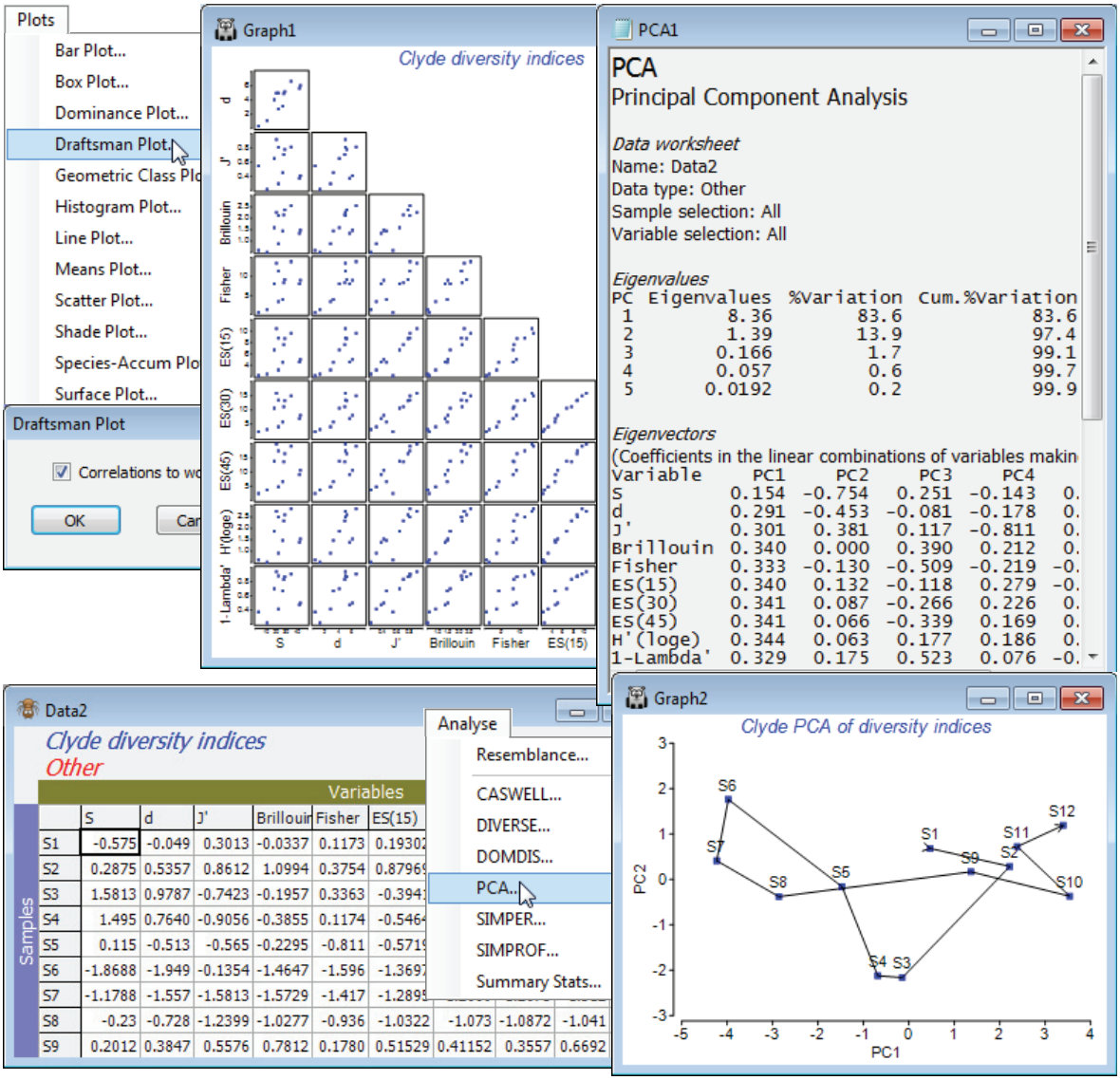
Rarefaction (Sanders/Hurlbert)

ES_n, the 'expected' number of species from n individuals ($n \le N$)

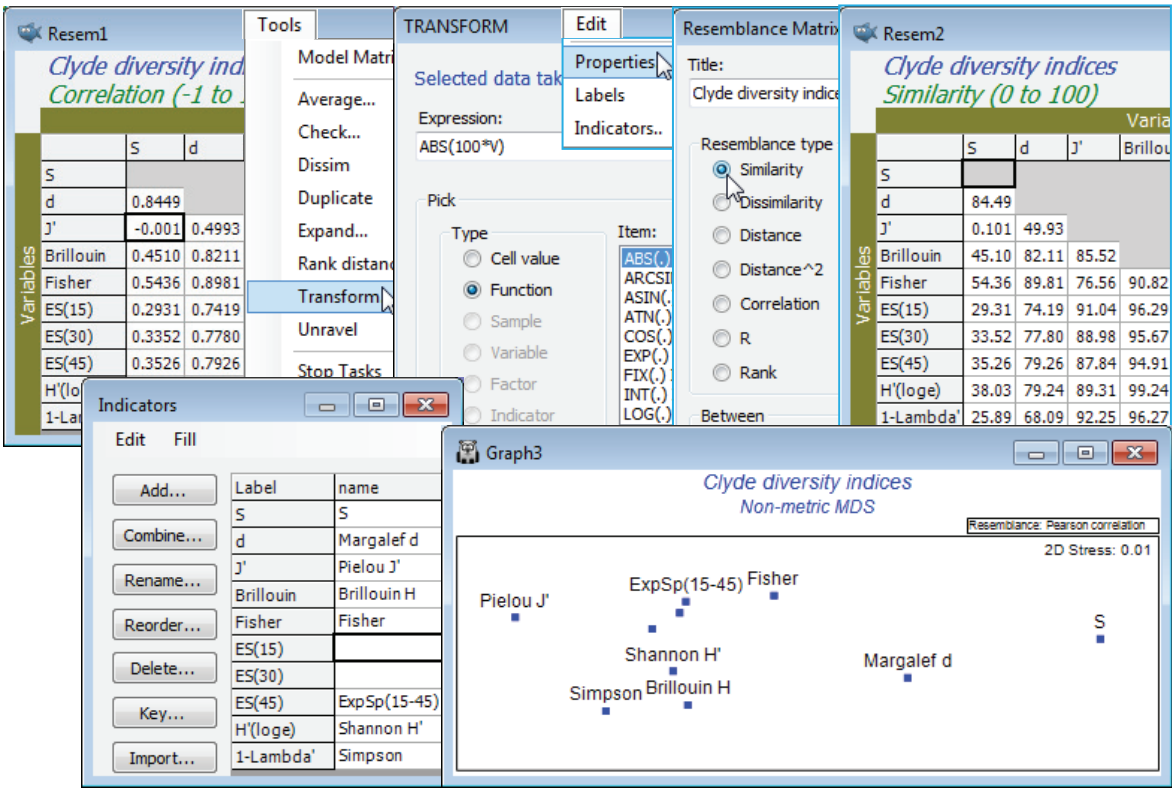


Multivariate analysis of diversities

For the diversity (variables) by samples matrix, Data1, **Plots>Draftsman Plot>**(✓Correlations to worksheet) shows that none of the indices is badly behaved, i.e. skewed, dominated by outliers, strongly curvilinear relationships etc, so no transforms seem called for. [To get the plot below, you might find it helpful to increase the symbol size on the **Samp. labels & symbols** tab, and on the **X & Y axis** tabs increase the title font sizes, unchecking (✓Limit size)]. Data1 needs **Pre-treatment> Normalise Variables**, however, before entry to **Analyse>PCA** since the indices are on different scales. On the configuration plot from PCA, turn off (✓Overlay vectors) on **Special>Overlays** and instead (✓Overlay trajectory) of the transect Site#. Site 6 is the dumpground centre, with Sites 1 and 12 at the extremities of the transect, and this combined set of diversity indices clearly displays the strong, simple gradient of effect, in a rather similar way to the full multivariate analysis of the original species data (you might like to carry out the latter, with a fairly severe transformation and Bray-Curtis similarities). The agreement is a consequence of the severity of the impact. The *meta-analysis* of Chapter 15 of CiMC shows this to be the most severe of the contaminant studies looked at there, but Chapter 14 also shows that such agreement is untypical, diversity measures being less likely to detect biological change for more intermediate-level disturbances. The PCA results (the eigenvalues) also make it clear that rather little is to be gained by calculating ten diversity indices instead of two or three: over 83% of the total variation in the 10 indices is accounted for by the first PC, and 97% (i.e. all of it, in effect) by the first two PC's. The coefficients (eigenvectors) show that the simple left to right gradient in the main axis (PC1) of the PCA is a roughly equally weighted combination of all measures (evenness + richness), both increasing away from the dumpground, whereas the second axis strongly contrasts the two main diversity components: PC2 is effectively (evenness – richness). This simplicity should not be a surprise, given the high correlations between indices evident from the draftsman plot, and from the correlation matrix Resem1 created with it.



A final, revealing plot can be produced from Resem1, by ordinating the variables. Technically, it first needs transforming before it can be considered a similarity matrix: there is a small, negative correlation between *S* and *J'*. It is effectively zero here, but other situations might produce large negative correlations, e.g. between equitability and dominance measures, and they should also imply similarity (of variables). **Tools>Transform>**(Expression: 100*ABS(V)) on Resem1 will achieve the conversion to a similarity matrix (and you could change its type on **Edit>Properties**). Then **Analyse>MDS>Non-metric MDS (nMDS)** generates the ordination plot for the variables shown below, in which the relative distances apart of the indices exactly reflects the rank order of their pairwise correlations (note that the MDS stress is effectively zero). The plot is largely linear, the extremities corresponding to pure richness (*S*) and evenness (*J'*), with other measures being a mix of these two components. The points have been more descriptively labelled using **Var. labels & symbols>**(Labels✓By indicator)>**Edit**, which is equivalent to **Edit>Indicators** on the Resem1 sheet, then **Add** an indicator: name. The boundary of the *n*MDS plot has also been appropriately reshaped for this linear plot, with **Special>Main>**(Plot type•2D>Aspect ratio: 3). Values of $n = 15$, 30 and 45 were chosen for the rarefaction indices *ES*(*n*) because larger values are not permissible, the site with lowest abundance having only 46 individuals. (To see this **Analyse>Summary Stats>**(For•Samples)>(✓Sum) on Clyde macrofauna counts, or just ask for ✓N in **Analyse>DIVERSE**). The fact that the *expected species numbers ES(n)* are clearly considerably closer to being evenness measures than the richness indices that their name implies (correlations of about 0.9 with *J'* and 0.98 with *H'*, compared with about 0.3 with *S*) results from the lack of ecological realism in their underpinning model. This assumes that individuals arrive randomly and independently into the sample, and hence the process can be reversed in rarefaction, by randomly excluding them. This does not correspond to the reality of a clumped spatial distribution seen for many species (as seen in Dispersion Weighting, Section 4). Resave the workspace Clyde ws2 for later use, and close it.



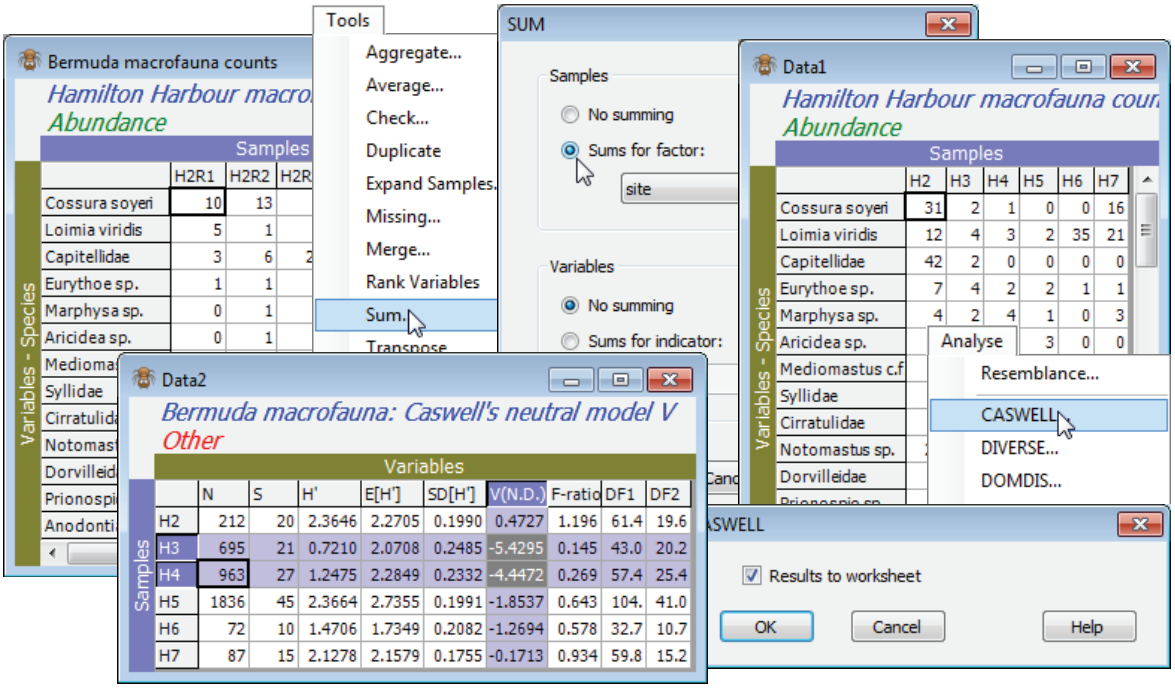
(Bermuda macrofauna)

Soft-sediment macrofaunal assemblages (along with meiofauna and biomarker suites) were studied at 6 sites in Hamilton Harbour, Bermuda (labelled H2, H3, H4, H5, H6, H7) during an international IOC workshop on the effects of pollutants in sub-tropical waters (Addison RF & Clarke KR, eds 1990, *J Exp Mar Biol Ecol* 138). There were 4 replicates at each site, giving a data matrix of 24 samples from 64 species, in the data file Bermuda macrofauna counts in directory C:\Examples v7\Bermuda benthos. These data will be used to illustrate computation of another diversity index, not now widely used (the validity of its assumptions being questionable for most assemblages) but which has been available in PRIMER from early versions and therefore retained for consistency.

Caswell's neutral model

Analyse>CASWELL generates *V* statistics for the Caswell neutral model, and is discussed in Chapter 8 of CiMC. It is essentially a comparison of Shannon diversity *H'* with the value it would be expected to take, conditional on the observed number of species *S* and individuals *N*, under some simple model assembly rules for the community, which are *ecologically neutral*, in the sense defined by Caswell H 1976, *Ecol Monogr* 46: 327-354. The normalised form of *H'* (subtract the modelled mean and divide by the modelled standard deviation) is the *V* statistic, positive values of *V* implying greater diversity than neutrality and negative values lesser. (There is an F test of its departure from $V = 0$, though this is not very convincing because it also depends on the neutral model assumptions, which are unrealistic for typical assemblages). The algorithm implemented here is due to Goldman N & Lambshead PJD 1989, *Mar Ecol Prog Ser* 50: 255-261.

Recreate the Caswell example in Chapter 8 of CiMC, for the Bermuda macrofauna counts by firstly summing across the replicates, to increase the sample size, with **Tools>Sum>**(Samples•Sums for factor: site) & (Variables•No summing). This is justified because there is equal replication at each site – **Tools>Average** would not be appropriate for a Caswell calculation because the entries are no longer real (integer) counts. Note that *V* could alternatively be calculated for each replicate, as for the diversity measures above, and this would allow standard means and confidence intervals based on variance estimates from replication, rather than the (less robust) internal variance estimate from the neutral model. On the summed Data1 take **Analyse>CASWELL>**(✓Results to worksheet), and the *V* values for each site (and the accompanying test calculations) are found in the resulting Data2 sheet, which can be manipulated, saved etc as with any other data matrix. Sites H3 and H4 are seen to have *H'* well below expectation under the neutral model (*V* statistics of -5.4, -4.5 respectively). Close the workspace – it will not be needed again.



Range of relatedness indices calculated

In order to obtain a diversity measure which steps outside the species abundance distribution, and which could therefore potentially strike out along a different axis to the linear richness-evenness combinations shown in the MDS of the mechanistic correlations among standard diversity indices, it would be helpful to introduce further attributes of the assemblage composition. One possibility is to combine biomass and abundance data, as in ABC curves (Section 16), but another – which we shall turn to now – is to introduce information on the relatedness of the species in each sample, as discussed at the start of this section. These indices are accessed through the final two tabs of the dialog box from **Analyse>DIVERSE**, namely **Taxdisc** and **Phylogenetic**. The nomenclature comes from the original papers on these topics (Warwick and Clarke's *taxonomic diversity* and *taxonomic distinctness* indices, and Faith's *phylogenetic diversity*), and does not imply that either set of indices is more appropriate to taxonomic or phylogenetic hierarchies. Other hierarchies (e.g. genetic, functional) could be equally appropriate and PRIMER does not now even need a hierarchy to compute the taxonomic distinctness measures – a *distance among species* matrix will suffice.

v7

The relatedness indices are all denoted by upper case Greek symbols, with superscript$^+$ if calculated from species lists. For definitions, and extensive discussion, see Chapter 17 of the CiMC manual.

Taxonomic distinctness

Quantitative:

Taxonomic diversity: Δ

Taxonomic distinctness: Δ^*

Presence/absence:

Average taxonomic distinctness (AvTD): Δ^+

Total taxonomic distinctness (TTD): $S\Delta^+$

Variation in taxonomic distinctness (VarTD): Λ^+

Phylogenetic diversity

Presence/absence:

Average phylogenetic diversity (AvPD): Φ^+

(Total) phylogenetic diversity: $S\Phi^+$ (Faith's 'PD')

Species distance information

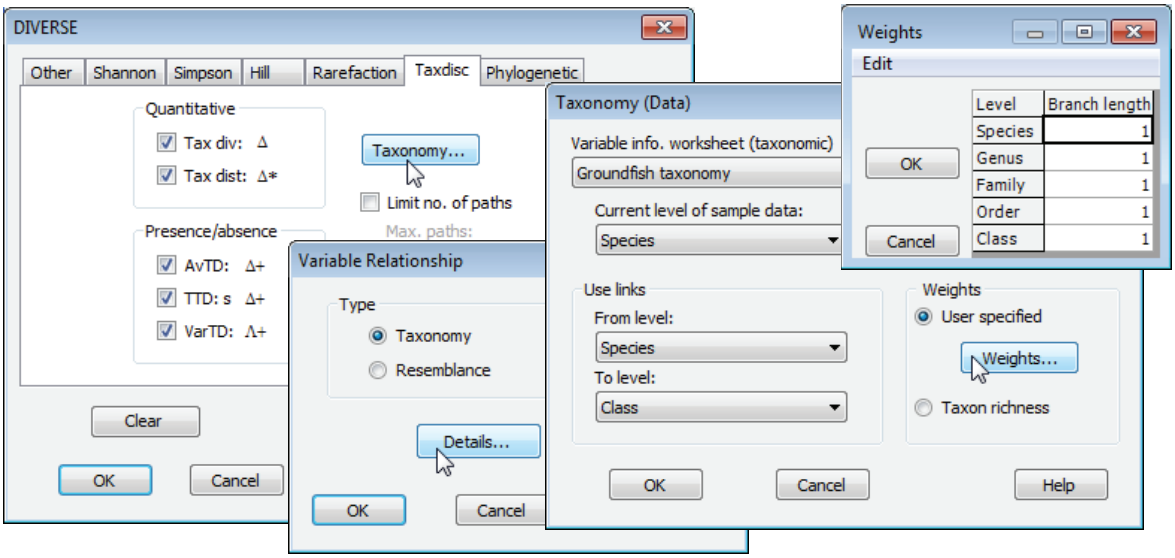
v7

For the first set of measures (on the **Taxdisc** tab), the **Taxonomy** button gives a choice of whether the *distances* among species (or whatever the variables represent) are provided by a tree structure (•Taxonomy) or a direct distance matrix among species (•Resemblance). The latter then requires a *Variable resemblance* matrix to be specified (perhaps one calculated among species on the basis of their traits, if this is to be a functional rather than taxonomic-based distinctness index). The former requires a *Variable information* sheet – usually an aggregation file of the type seen near the start of Section 11 – which needs to be in the workspace before **Analyse>DIVERSE** is run (if only one such file has been read in, it will be the default). This is a look-up table which gives a taxonomic (or other) tree of all species, allowing the routine to calculate species distances internally (these are not actually output but could be so, if needed, by **Analyse>Similarity** when the active window is the aggregation worksheet, as seen in Section 5). For the second set of measures (the **Phylogenetic** tab in the DIVERSE dialog), the **Taxonomy** button offers only the option to input a *Variable info.* worksheet because the PD measures (Φ^+ and $S\Phi^+$) can only be computed from a species tree and not from a triangular matrix of between-species distances.

v7

Distances in aggregation worksheets

Such tree structures (e.g. taxonomies) are one of a distinct worksheet type, *Variable Information*, slightly expanded in PRIMER 7 from the aggregation file format of PRIMER 6, but still with an *.agg extension when saved as PRIMER 7 binary format – they can also be input or output in *.xls or *.xlsx Excel format. The aggregation matrix could simply be a tree constructed for just those species in the current data matrix or it could be a wider and more comprehensive *master list* for those faunal groups. The species (or other variable) labels used in the data worksheet must find an exact match in the labels of the aggregation sheet (or, if working from a higher taxonomic level in the aggregation matrix, e.g. genus, used as the variable names for the data sheet, then this must be specified in *Current level of sample data*). The species do not need to occur in the same order in the both sheets because of PRIMER's use of strict label matching. See Section 11 for information on checking aggregation arrays for inconsistency – potential mis-spellings – with **Tools>Check**.



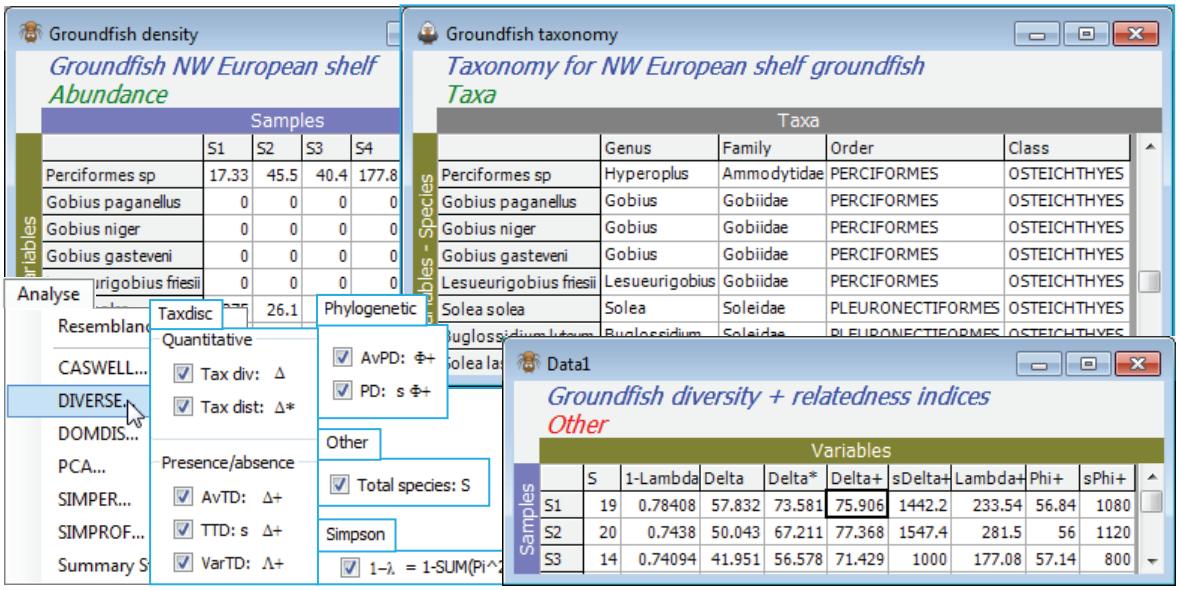
There are also options under the Taxonomy (Data) dialog to use only part of the taxonomic tree. For example, (Use links)>(From level: Genus) would start from genus level – in effect treating all species in the same genus as the same taxon – which is not often a requirement but could be useful if the identifications are very patchy to the species level, but reliable to genus. Similarly, the tree could be compressed at the top level so that, for example, no greater distance is assumed between two species in different classes than for two species in different orders but the same class – that would be achieved by specifying (Use links)>(To level: Order).

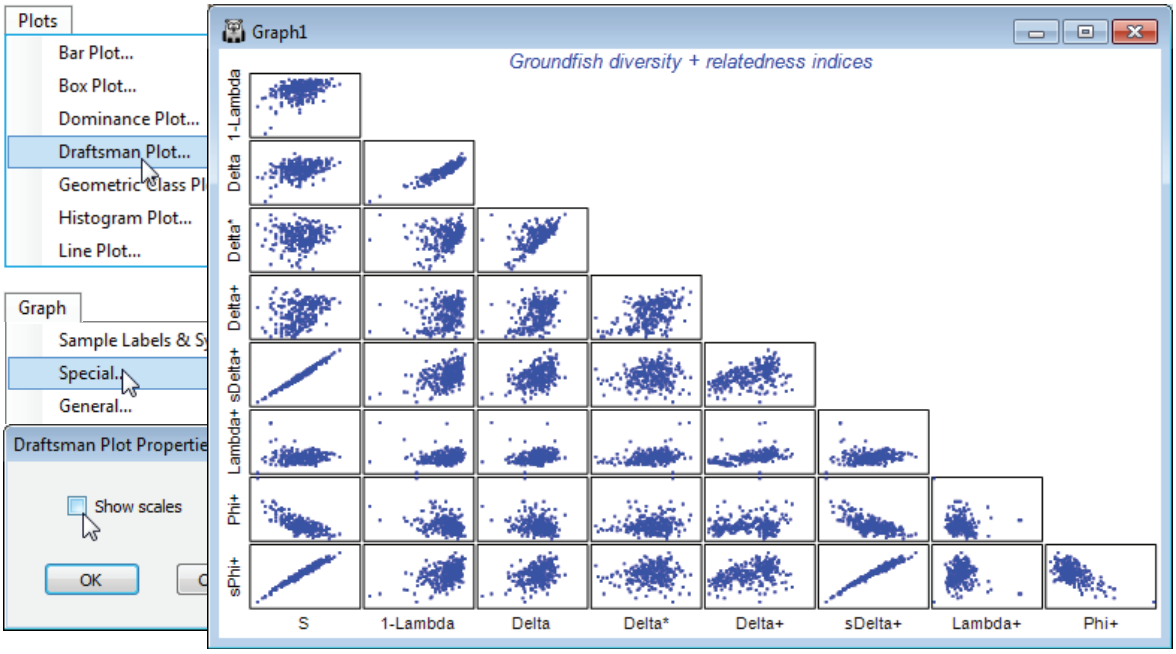
Weighting of tree step lengths

The other box in this Taxonomy (data) dialog can be used to alter the weights given to the various branch lengths in the tree (and includes the previous compression at the top or bottom of the tree as a special case, with those step lengths set to zero). By taking (Weights•User specified)>**Weights**, the default lengths are displayed: equal steps are assumed, and any values placed here will always be standardised, subsequently (and automatically), so that the longest path in the tree is set to 100. Thus a change to step lengths of 2 for all categories would not alter the values of any of indices, but a change to decreasing step lengths of 6 (species to genus), 5 (genus to family), 4 (family to order) etc. could be worth exploring because it would put relatively more weight on the shorter branch lengths between species (of which there are fewer) rather than leaving much of the emphasis on the longer branch lengths (because there are many). One logical basis for altering the step lengths from their default would be to make them depend on the decrease in the number of taxa in the master list when making that step – the smaller the decrease in the number of taxa, the shorter the step length. This has the merit of consistency if, for example, an arbitrary taxonomic level (e.g. subfamily) is interpolated but not used (i.e. there are as many subfamilies as families in the master list). The set of distinctness indices would then remain unchanged. The detail is given in Clarke KR & Warwick RM 1999, *Mar Ecol Prog Ser* 184: 21-29, and their weighting scheme can be implemented here by taking (Weights•Taxon richness) in the Taxonomy (Data) dialog box.

Taxonomic distinctness (European groundfish)

The aggregation matrix for the NW European beam-trawl survey data on groundfish assemblages (93 species in 277 samples, from 9 sea areas) was last seen in Section 11, where it was checked for consistency. However, the workspace is now rather cluttered so open a new one in C:\Examples v7 \Europe groundfish, containing data Groundfish density and Groundfish taxonomy, and save it as Groundfish ws2. Here, data and aggregation matrices have the same full set of species, in the same order. With Groundfish density as active sheet, run **Analyse>DIVERSE** and on the **Taxdisc** and **Phylogenetic** tabs, check (✓) all the quantitative and presence/absence options: Δ (= delta), Δ*, Δ^+, $S\Delta^+$, Λ^+ (= lambda+), Φ^+ (= phi+) and $S\Phi^+$, taking also (✓Results to worksheet). Under **Taxonomy** >(Type•Taxonomy) take all the defaults: (Variable info. worksheet: Groundfish taxonomy) & (Current level of sample data: Species) & (Use links>(From level: Species) & (To level: Class)) & (Weights•User specified), with the **Weights** left on their values of step lengths of 1 between all levels. Take also the number of species (*S*) and Simpson evenness 1-λ, from the **Other** tab. Look at the correlation between these indices by **Plots>Draftsman Plot**. (To obtain the plot overleaf, the axis scales have been switched off by unchecking (✓Show scales) from **Graph>Special**).





In the draftsman plot, note particularly the first column of plots, which set each index against the number of species, *S*. These bear out the general observations of Clarke KR & Warwick RM 2001, *Mar Ecol Prog Ser* 216: 265-278, and Chapter 17 of the CiMC manual, that:

a) total phylogenetic diversity PD ($S\Phi^+$) and total taxonomic distinctness TTD ($S\Delta^+$) are dominated by *S* (which will be strongly influenced by the differing sampling effort for the 277 rectangles);

b) an attempt to correct for this by using average PD (Φ^+) is unsuccessful, there still being a strong correlation with *S* (negative now), but it is successful for average taxonomic distinctness AvTD (Δ^+) and variation in taxonomic distinctness VarTD (Λ^+), Clarke & Warwick 2001 showing that (mechanistic) independence of Δ^+ and Λ^+ from *S* is to be expected on theoretical grounds;

c) quantitative taxonomic diversity (Δ) retains a strong element of the evenness component from the species abundance distribution, i.e. is strongly correlated with Simpson's $1–\lambda$. In fact, Δ is a compounding of Simpson's $1–\lambda$ and a pure relatedness index, thus quantitative taxonomic distinctness $\Delta^* = \Delta/(1–\lambda)$ more nearly represents pure relatedness, and is seen to be much less positively correlated with evenness (here as Simpson $1–\lambda$ but the same is true for Pielou's *J′*, or even Shannon *H′* – which is largely an evenness measure, with a small component of *S*);

d) the quantitative (Δ^*) and pres/abs (Δ^+) forms of AvTD, though positively correlated (≈ 0.5), are not highly so, suggesting (as other evidence does) that they capture somewhat different aspects of relatedness and are both worth examining when quantitative data exists;

e) because of their use of the taxonomic tree structure, the taxonomic distinctness measures capture an axis of variation in the samples not reflected by the standard diversity measures (this can be seen by repeating the PCA, and the MDS variables ordination, near the start of this section, for the above relatedness indices together with the classic measures *S*, *d*, *J′*, *H*, α, *H′* and $1-\lambda'$).

Box plots & means plots for diversity indices

v7

The sheet Data1 of this suite of diversity indices for each of the 277 samples, split into 9 sea areas (factor area), could now be input to two new multi-plot routines in PRIMER 7, namely standard univariate box plots and means plots, treating the sea areas (1: Bristol Channel, …, 9: E Central North Sea; see map Fig. 17.10 in CiMC) as a group structure, with an average of about 30 replicate sample boxes (quarter degree rectangles) within each sea area. Taking **Plots>Box Plot>**(Group factor: area) on Data1 gives 9 separate box plots, *Graph2* to *Graph10*, one for each diversity index in the above set, each with 9 'box and whiskers' constructions, one for each area. These are placed into a multi-plot, *MultiPlot1*, and are intended as 'quick look' plots, with limited flexibility for manipulation (individual plots restricted to choice of axis scales, title content and text sizes). For Data1 again, **Plots>Means Plot>**(Group factor: area) & (✓Join means) & (✓Common variance estimate) gives a similar set of 9 plots within *MultiPlot2*, each of observed means and confidence intervals for the true mean of that particular diversity index for each of the 9 areas. There is choice of separate variance estimates for each area, or a common variance estimate (as from the ANOVA residual mean square). Interval widths for means vary here because areas have differing replication.

Testing taxonomic distinctness against a master list

Wide-ranging biogeographic studies, and particularly historic data, are often restricted to simple species lists. Even where quantitative information exists, it is rarely from sampling protocols that have been standardised with respect to sampling effort over the whole data. Where sampling is so exhaustive that the asymptote of the species-area curve is approached, then it may be valid to compare diversity status by the length of these lists (species richness *S*), but this is not often the case (in marine science, certainly). As is well known, *S* is heavily sampling effort dependent so, if sampling effort is variable and unknown, any valid statements about richness appear problematic. However, the two relatedness measures discussed earlier, average taxonomic distinctness (AvTD, Δ^+) and variation in taxonomic distinctness (VarTD, Λ^+), can not only be calculated from simple species lists, with the added knowledge of their Linnaean (or other) classification, but also possess a robustness to the varying number of species *S* in the lists. To be more precise, in different-sized sublists generated by random sampling from a larger list (simulating the action of sampling with variable effort) their mean values are unchanged. This suggests that it is valid to compare Δ^+ (or Λ^+) over historic time or biogeographic space scales, under conditions of variable sampling effort. (Note that the indices are *average* not *total* measures, and orthogonal to species richness – along a third PC diversity axis, would be one way of thinking of it – and therefore an *addition* to *S*, rather than a substitute for it, in cases where sampling effort is controlled and *S* can be validly compared.)

Furthermore, a test can be constructed for the null hypothesis that a species list from one locality (or time) has the same taxonomic distinctness structure as a 'master' list (e.g. of all species in that

biogeographic region) from which it is drawn. This is again by simple randomisation: given there are s species observed in a particular sample, make repeated drawings at random of s species from the master list and compute Δ^+ for each drawing, building up a histogram and a 95% probability range of values of Δ^+ expected under the null hypothesis, with which the true Δ^+ can be compared. Values below the lower probability limit suggest a biodiversity that is 'below expectation'. This can be carried out for a range of sublist sizes and the limits plotted against s, to give a 95% *funnel plot* of expected values (the funnel arises from uncertainty being greater for smaller sublists). This can be repeated for VarTD (Λ^+), giving a second set of histograms and funnel. Together, the true Δ^+ and Λ^+, and the simulated values obtained by drawing their number of species from the master list, can be plotted on a single (x,y) scatter plot. Probability regions ('egg-shaped' contours, called *ellipse plots* since they are back-transformed ellipses) covering 95% of the simulated values can be drawn for a range of sample sizes, and the true (Δ^+, Λ^+) compared with their appropriate contour.

TAXDTEST (European groundfish)

v7

Further theoretical details and discussion can be found in Chapter 17 of CiMC, which also presents analyses for the Europe groundfish data, whose workspace Groundfish ws2 should still be open. These taxonomic distinctness tests (on presence/absence data only) are accessed by **Analyse> TAXDTEST** either when the active window is either a variable information sheet (an aggregation file) or a variable resemblance matrix. These determine the *master list* (**Master taxonomy** on the TAXDTEST dialog box) from which random subsets of species will be drawn, in order to construct the probability histogram, funnel or ellipse plots. It is also the default aggregation sheet used in calculating the observed Δ^+ and Λ^+ for any specific set of samples, to superimpose as points on the simulated funnels or ellipses (Sample data✓Use Sample data>Taxonomy•Use master). However, with (Taxonomy•Specify different>**Taxonomy**), a different aggregation sheet could be supplied, for the sample data calculation only. This would normally be quite unnecessary because the species relatedness needed for any particular sample can be drawn from the master taxonomy: as noted earlier, there is no necessity for the sample data matrix to contain all the same species in the same order as the aggregation (or variable resemblance) sheet – it is just necessary that all the species are found in the master list. However, it could be valid to place data from a region (or geological time), with its own aggregation information, on an expected funnel from an entirely different region (or time), with a different master list, so this option is catered for. If based on a variable information sheet (aggregation file), **Taxonomy** buttons will give the dialog seen earlier, allowing compression of the taxonomic tree and path step lengths which can be altered from equal weighting.

Compute time & limits on path numbers

v7

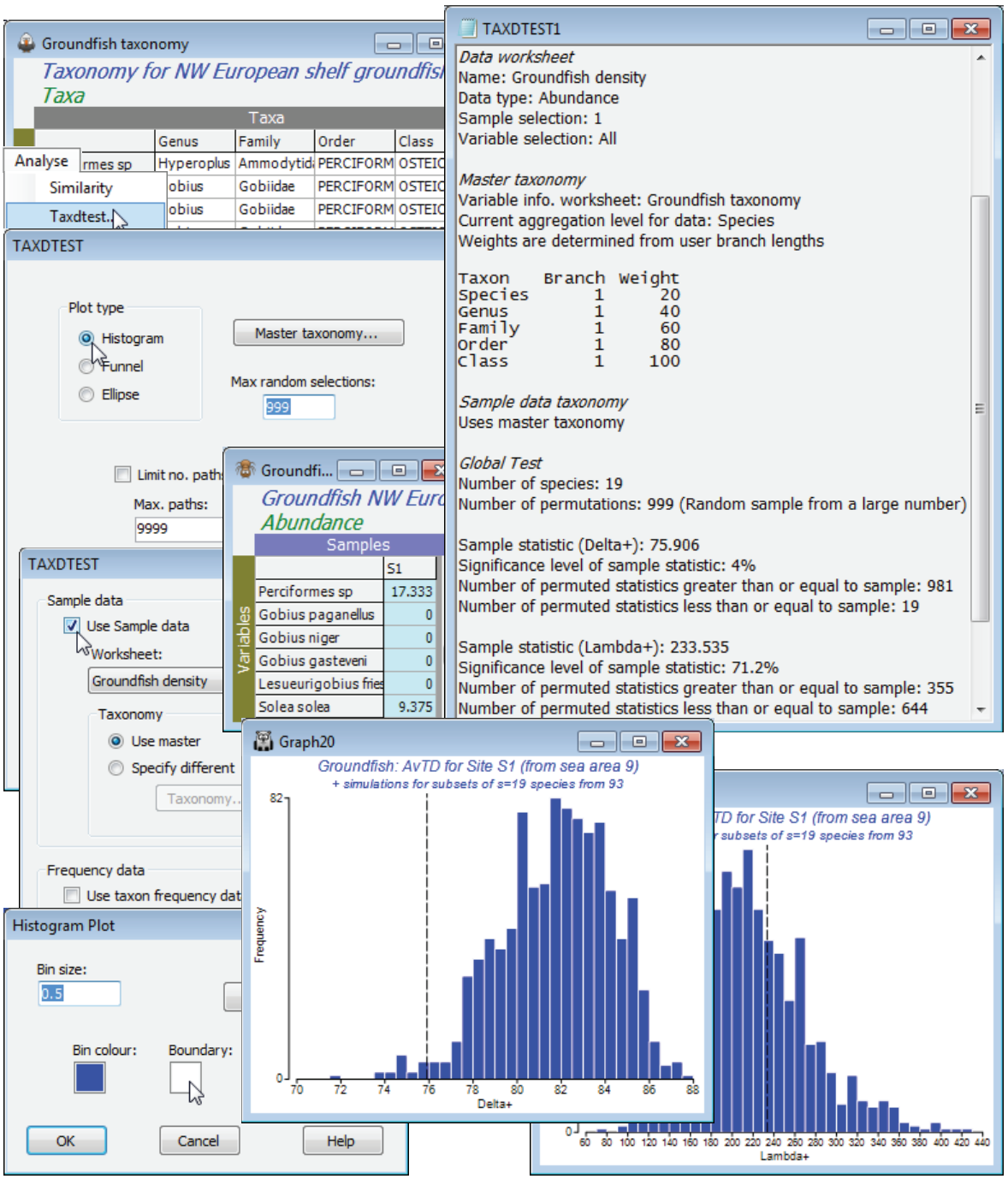
A new option in PRIMER 7 recognises that computation time can become an issue for particular relatedness analyses when the master list is extremely large - as could happen if, for example, the world list of fish species, or the entire marine species directory of European waters is input as the master list (a species list of 10,000 has 100 million path lengths between all pairs of species). But it is not necessary to calculate all of these to know the true AvTD Δ^+ of the master list, for example – we can again exploit the unbiasedness of random samples to get all the accuracy we need without complete computation, and this option is taken with (✓Limit no. paths)>(Max paths: 9999), say. This option was also provided for distinctness estimates in the DIVERSE routine but may be less necessary there and is inappropriate, so should be avoided, for the quantitative Δ, $\Delta*$ calculations. Path limitations are not the default and are best saved for use only when essential to obtain results.

Histograms for one sublist size

For an example, take the first of the 277 groundfish samples, the 0.25° rectangle S1. Highlight and select just this column from Groundfish density, with **Select>Highlighted** (this is a quantitative matrix not presence/absence, but TAXDTEST will automatically convert it to P/A data – as does DIVERSE when computing Δ^+, Λ^+ etc). With the Groundfish taxonomy sheet as active window, run **Analyse>TAXDTEST>**(Plot type•Histogram) & (Max random selections: 999), with defaults for the **Master taxonomy** button and, on the next screen, check (Sample data✓Use Sample data)> (Worksheet: Groundfish density)>(Taxonomy• Use master). Leave the (Frequency data) section for now – it will be demonstrated later. The routine counts S = 19 species in the supplied sample data column so produces 1000 random draws of 19 species from the master list, Groundfish taxonomy. It then calculates Δ^+ and Λ^+ for each random draw and puts the values into a histogram for each index. The real values of Δ^+ and Λ^+ for that data column are shown by a dashed line, as usual, and the significance levels (here, for a two-sided test) are given in the results window. In this case, only 19 of the 999 random draws gave Δ^+ values less than or equal to the real Δ^+. The probability of this

under the null hypothesis (that species at that S1 location are representative of the full taxonomic spread in the master list of 93, so retain the overall biodiversity) is ≤ (19+1)/(999+1) = 0.02, i.e. a significance level of ≤ 2% on a one-sided test. It is arguable here that the test should be one-sided, and that the only departure of interest from the null is one of decreasing taxonomic distinctness – perhaps through extensive beam-trawling differentially affecting groups of groundfish higher taxa with particular life-history characteristics. There may, however, be situations in which we would like also to be able to detect increases in Δ^+, and it is certainly true for Λ^+ that plausible alternatives to the null hypothesis could be two-sided. So PRIMER quotes two-sided significance levels in both cases (thus a significance of 4.0% for Δ^+) – a one-sided test would simply halve the quoted values. Also remember that each run will give slightly different results because of different random draws, and in borderline cases you might want to increase the number of random draws, e.g. to 9999.

v7 The histograms are displayed in a multiplot, with just the two component plots. The usual display options are accessed through the **Graph>General** menu, to change overall font size, titles etc, and **Graph>Special** has here allowed the bin size to be increased for a smoother histogram, and can *v7* allow colour change of the histogram bars and boundary (the latter from black to white here).

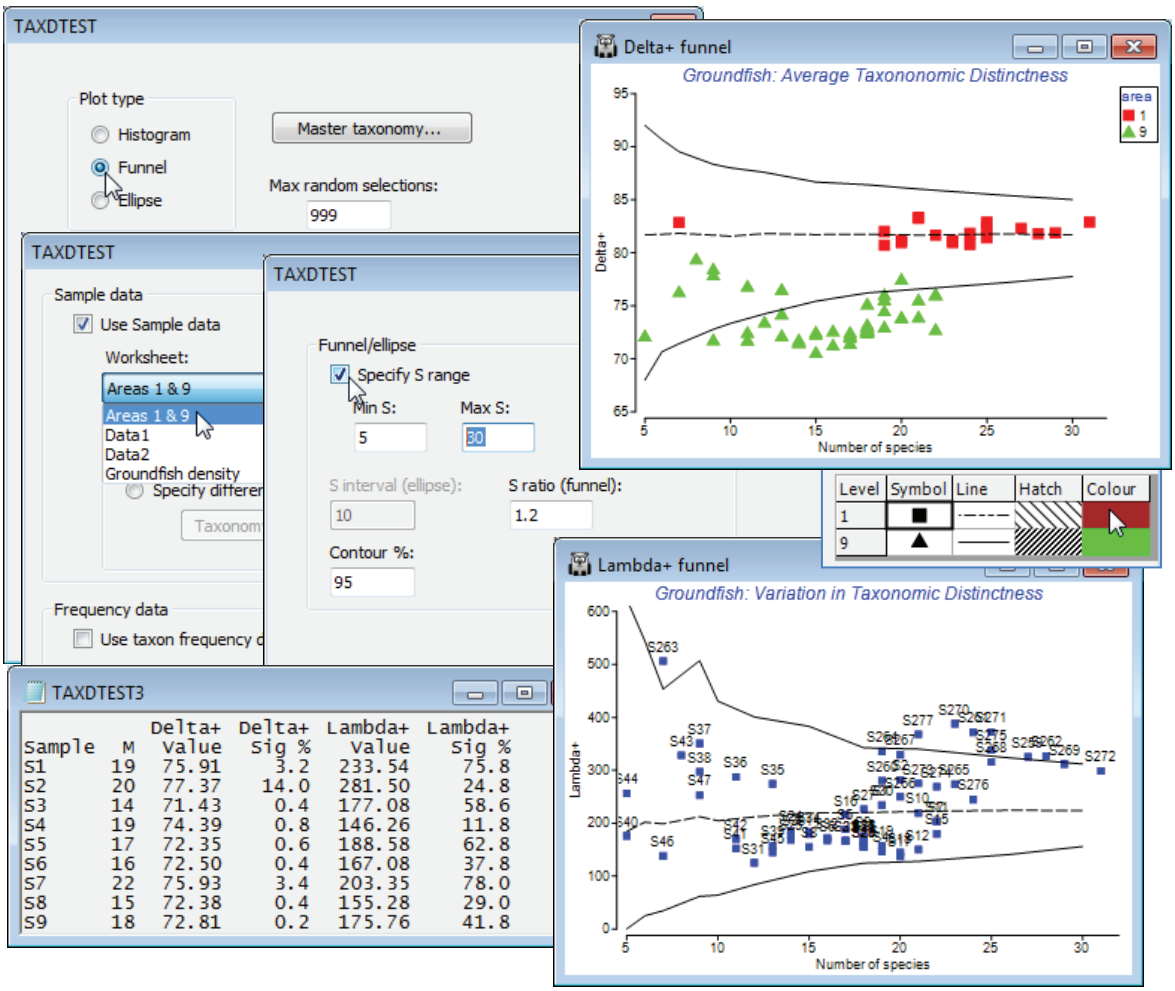


If you submit several columns of data by mistake at this stage, the error message *Only one sample must be selected for histogram* will result. If you wish to generate histograms of expected Δ^+ (or Λ^+) values from the master list, for a fixed sample size (e.g. $S = 20$), without referring to a specific data sample, then uncheck (✓Use Sample data) in the TAXDTEST dialog. You will then be asked to supply that size, e.g. Histogram>S value (no sample data): 20.

Funnels for a range of sublist sizes

It is impractical to produce detailed histograms, such as those above, for each of the 277 samples, so a preferable option is just to view the 95% lower and upper limits for a range of sample sizes *S*, using a funnel plot so that a set of samples can be plotted on this. So, first select all sea area 9 (E Central N Sea) and sea area 1 (Bristol Channel) samples from Groundfish density, with **Select> Samples>**(•Factor levels)>Factor name: area>**Levels**, leaving only 1 and 9 in the Include box, and **Tools>Duplicate** this, renaming it Areas 1 & 9 (and remove the selection on the original sheet with **Select>All**, for later use). Then run **Analyse>TAXDTEST** again, on Groundfish taxonomy, with (Plot type•Funnel) & (Max. random selections: 999) and **Next**>(✓Use Sample data>Worksheet: Areas 1 & 9). Now, **Next**>(Funnel/ellipse✓Specify S range)>(Min S: 5) & (Max S: 30), to span the spread of *S* values on the display. The (S ratio (funnel): 1.2) option determines how many *S* values are calculated in the range 5 to 30, the *S* values stepping up by multiples of 1.2 by default (then rounded), thus $S = 5$, then 6 (=5×1.2) etc. The final box on this screen gives 95% intervals if the default is taken (2.5% of simulations fall above the upper limit and 2.5% below the lower limit).

The results and funnel plots for Δ^+ and Λ^+ are shown below and indicate that, whilst area 1 samples are within expected ranges for average taxonomic distinctness, based on the 93 species master list, area 9 samples have reduced diversity (AvTD is the more easily interpretable of the two indices, since it measures the average breadth of the assemblage). Rogers *et al* 1999 (reference in Section 5) discuss possible reasons for this. Note that these plots have been tidied up, with **Graph>Sample Labels & Symbols**, by removing the labels and adding symbols for factor area, changing symbol size/colour etc, as for any other plot. The probability limits could be further smoothed by running with (Max random selections: 9999) but will still show kinks for small *S*, because *S* is discrete.

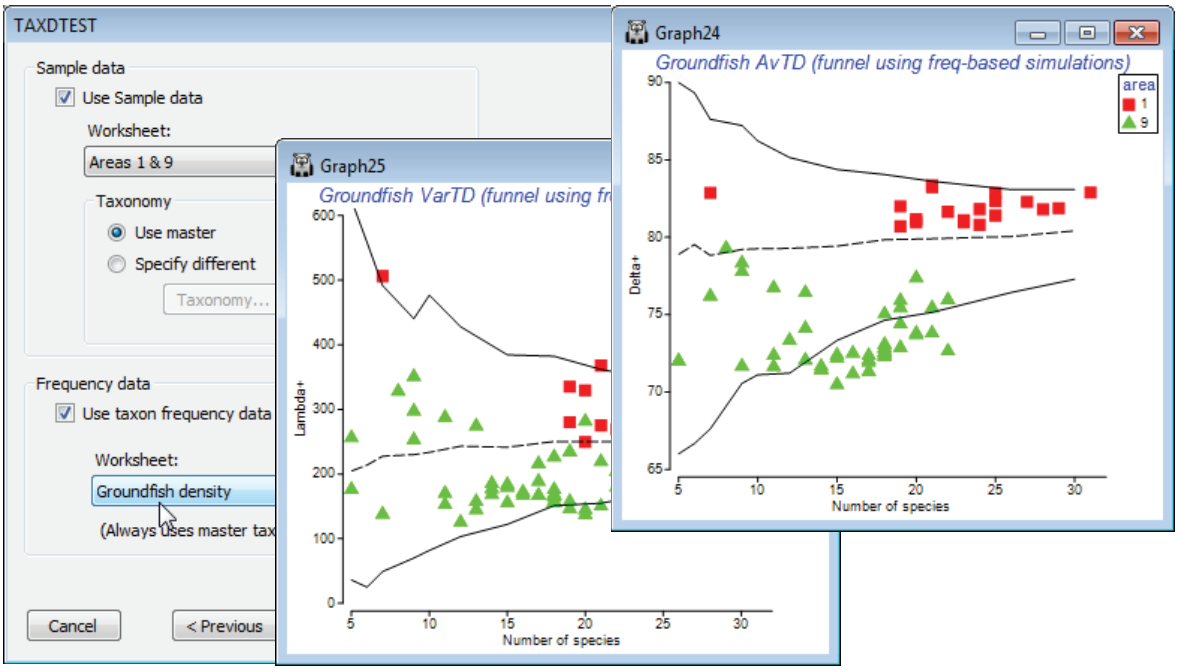


Using taxon frequency in simulations

Another option on the TAXDTEST dialogs is that the simulation of random draws from the master list, to generate histograms, funnels etc, can be constrained to match the probabilities of occurrence of each species, as observed in a large set of samples defining those taxon frequencies. Thus certain species are picked more often in the random subsets, because they are observed to be present more often in real samples of this type. The simulated mean and range of (e.g.) AvTD values generated in this way could be argued to give a more realistic yardstick for assessing the observed AvTD. These are produced by checking (✓Use taxon frequency data) and supplying a data matrix (which will be turned into P/A, if it is not already that), with a wide spread of samples of the full set of species in the master taxonomy, which can be used to calculate frequencies of occurrence.

An natural example here would use the full Groundfish density sheet (having removed the earlier selection), with its large number of samples (277) determining probabilities of occurrence of each of the 93 species in any one sample. Now run **Analyse>TAXDTEST** on Groundfish taxonomy, again with (Plot type•Funnel), the default taxonomy options and (✓Use Sample data)>(Worksheet: Area 1 & 9), as before, but with (✓Use taxon frequency data)>(Worksheet: Groundfish density). Specifying S ranges as previously produces the plot shown below, in which the frequency-based simulated mean is no longer exactly independent of the sub-list size *s*, though the increase with *s* is seen to be slight here, on the scale of the probability limits, and the conclusions would be largely the same. Of course, the real Δ^+ values are unchanged – they are not a function of assumptions made about the relevant master list to simulate from, or whether to carry out simple random or frequency-based simulations. And naturally, if your study does not lend itself to testing hypotheses about assembly rules of species drawn from any sort of regional master list, you can simply use the taxonomic indices in the same way as demonstrated earlier for a range of diversity measures, in a purely comparative way across a series of groups, in univariate means plots or ANOVA tests based on the replicate information. (E.g. you can select a single measure, such as Δ^+, and take Euclidean distances on its 277 values across all rectangles here, inputting that resemblance matrix to the PERMANOVA routine in the PERMANOVA+ add-on, to give exactly the ANOVA table for a one-way test of the area factor, with the F value tested by permutation, not F distribution tables).

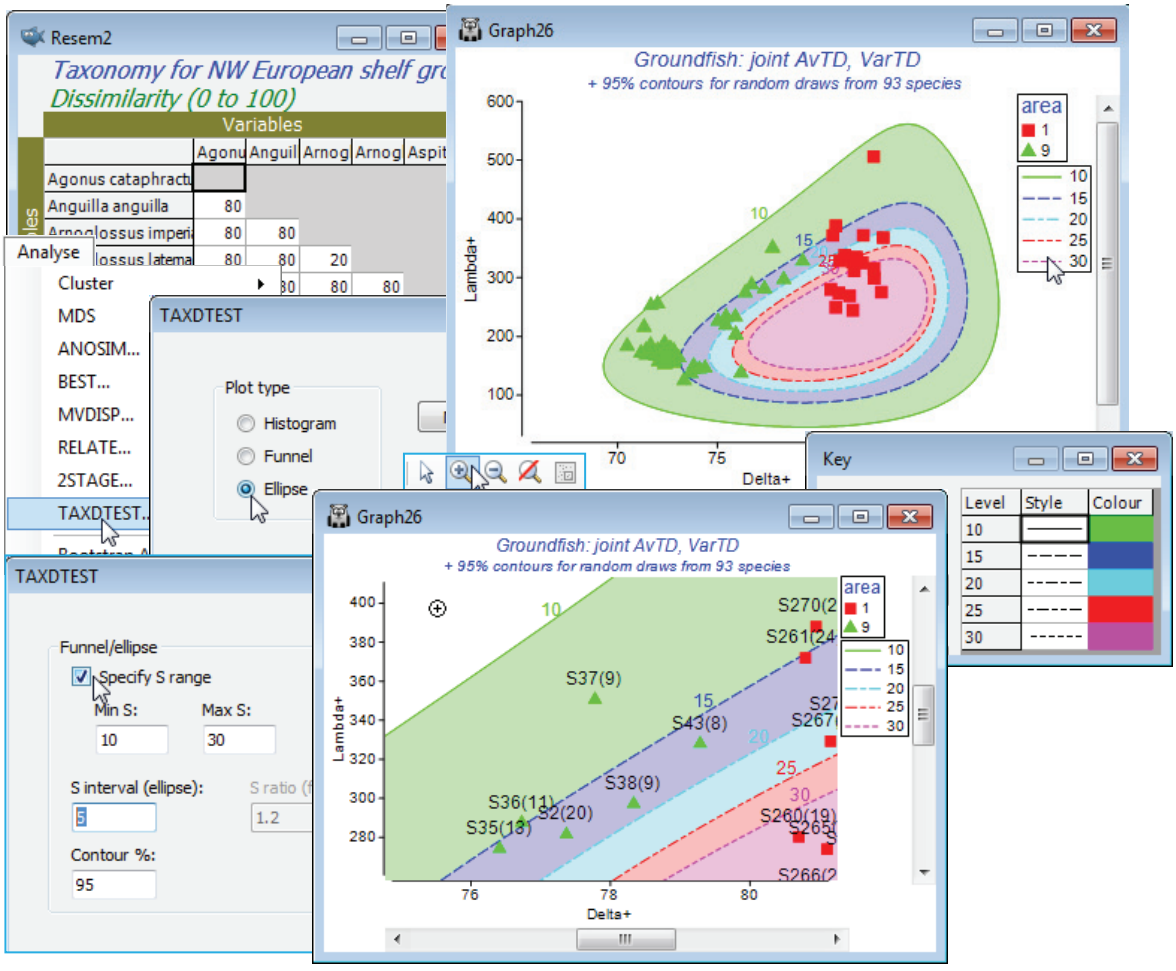
v7



'Ellipses' for joint values of (Δ^+, Λ^+)

The final option is to consider Δ^+ and Λ^+ in combination, by plotting 95% probability contours for their joint distribution, under the null hypothesis of simple random (or frequency-based) selection from the master species list. Optionally, pairs (Δ^+, Λ^+) from a real sample data matrix can be added. There may be some advantage in looking at both measures simultaneously because departures from expectation may reveal themselves as, say, lowish Δ^+ and highish Λ^+ values, neither of which was significant on its own, but in combination outside the joint (Δ^+, Λ^+) contours, for which Δ^+ and Λ^+ might be negatively correlated. (The contours are drawn by approximating the simulations by a bivariate normal distribution in a transformed space, then back-transforming – Chapter 17, CiMC).

v7 Just in order to create an example of how TAXDTEST can be run from a variable similarity matrix (such as might be found in a functional rather than taxonomic description of species relatedness, thus creating an Average Functional Distinctness diversity, AvFD, see Somerfield *et al* 2008, *ICES J Mar Sci* 65: 1462-1468), take **Analyse>Similarity**>(•Taxonomic), which simply returns a matrix of distances through the taxonomic tree. With this (Resem2) as the active window, run **Analyse> TAXDTEST** with option (Plot type•Ellipse) & (✓Use Sample data>Worksheet: Area 1 & 9) & (✓Specify S range)>(Min S: 10) & (Max S: 30) & (S interval (ellipse): 5), and selecting simple random sampling, i.e. uncheck (✓Use taxon frequency data). With (Contour %: 95), five contours will be produced, within which approximately 95% of the (Δ^+, Λ^+) pairs will lie, for s = 10, 15, 20, 25 and 30 random species draws. These contours must logically be concentric – if they do not look so it is certainly worth specifying more simulations, e.g. by (Max random selections: 9999) on the first TAXDTEST dialog screen. You may need to change the symbol types/colours again to get the first plot below, depending on which part of the Explorer tree you made the change to the **Key** area previously (if it was in the Area 1 & 9 sheet itself then this will be retained). There will now also be a key which controls the line type and line/shading colour for the 95% contours, and though this can be accessed from the **Keys** tab on the Graph Options dialog box, if changes are needed it is v7 simplest just to click on the line key in the plot itself, taking you into the colour dialog.



For each sample, the idea is to visually interpolate between the contours for the two s values that straddle its observed number of species S, and determine whether that point is inside or outside its expected 95% contour (a Bonferroni-type correction could be used for the probability limits, or you should just bear in mind in interpreting the plot that 1 in 20 of points will fall outside 95% limits under random draws!). The conclusion here is again of a lower than expected average taxonomic distinctness (but mid-range VarTD) for area 9, and this is discrete from area 1, which has expected mid-range AvTD (and little evidence of VarTD being higher than expected). The interpretation of Δ^+ and Λ^+ in general is covered in Clarke KR & Warwick RM 2001, *Mar Ecol Prog Ser* 216: 265-278 and Warwick RM & Clarke KR 2001, *Oceanog Mar Biol Ann Rev* 39: 207-231, and this study specifically in Rogers *et al* 1999, *J Anim Ecol* 68: 769-782 and Chapter 17 of CiMC.

16. Diversity curves (*Geometric Class*, *Dominance* and *Species-Accumulation Plots*)

Range of diversity curves

PRIMER plots a range of what might be termed *diversity curves*, under the **Plots>Geometric Class Plot** and **Dominance Plot** menus, obtained when the active window is a data matrix, and these are described in Chapter 8 of CiMC. They display a combination of evenness and richness components of diversity in a more continuous way than is achieved by a single index and, in the case of ABC (Abundance-Biomass Comparison) curves, incorporate both abundance and biomass components of the assemblage in a single plot. The multiple ABC plots (one for each sample, matched over the abundance and biomass arrays) are now held in a v7 multi-plot. Factor levels can be identified on the plots by symbols/line types and a structure for testing dominance plot differences over sites/times/treatments etc, facilitated by **Analyse>DOMDIS**, generates a triangular matrix of between-curve distances which can be input to **ANOSIM** (Section 9). Richness (*S*) estimators are provided in **Analyse>Species-Accum Plot**, which (as elsewhere) supplements the plot by sending statistics to a new worksheet, making it easy to export information from PRIMER to other software. (There are effectively no changes of functionality for the routines in this section, by comparison with v6).

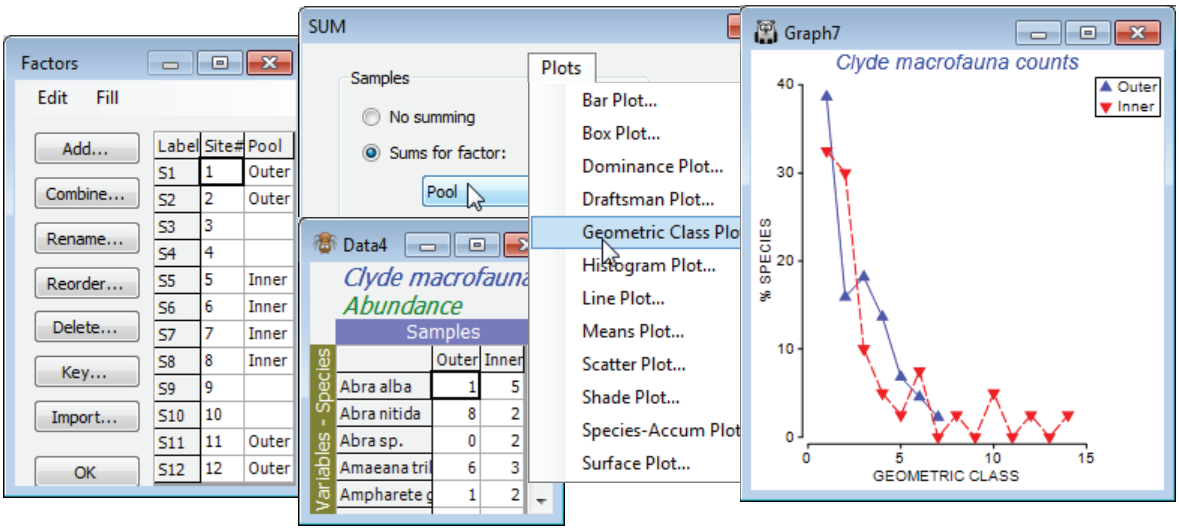
v7

Geometric class plots

These are essentially multiple frequency polygons, plotted on a single graph, for each sample in the active sheet, which needs to be a taxon (species) by samples array of genuine counts. If you wish to plot a single curve for each of a number of groups of samples then you should first pool replicates in each group with **Tools>Sum** – or, to pool all the samples in an array into a single sample, you can use **Analyse>Summary Stats>**(For•Variables)>(✓Sum). Then **Plots>Geometric Class Plot** gives several line plots (or just one) on a single (*x,y*) graph in which the *y* axis is the number of species that fall into a set of geometric (× 2) abundance classes (*x* axis). That is, each line on the plot gives the number (or %) of species represented in the sample by a single individual (class 1), 2-3 individuals (class 2), 4-7 individuals (class 3), 8-15 individuals etc. Statistical ecologists call these SAD curves (*Species Abundance Distribution*), and there is much early literature on fitting by distributions such as the truncated log-normal, proposed on (unconvincing) theoretical grounds. Fisher RA *et al* 1943, *J Anim Ecol* 12 was the first (as in so much else, statistically!) to model such data, fitting it to the single-parameter log series distribution – this parameter (α) is the Fisher index calculated by **Analyse>DIVERSE**, see the previous section. It has been suggested that impact on assemblages changes the characteristic form of the SAD curve, lengthening the right tail because some species become very abundant and other, rarer, species (singletons) disappear.

v7

Close the existing workspace (it is not needed again), and re-open the Clyde dumpground study, workspace Clyde ws2 analysed at the start of the last section, or just open the abundance matrix Clyde macrofauna counts from C:\Examples v7\ Clyde macrofauna. The plot would be cluttered with all 12 transect samples displayed, so contrast just two sets of pooled samples – the outer (1, 2, 11, 12) and inner (5, 6, 7, 8) sites – the pools need to be the same size for unbiased comparison. For Clyde macrofauna counts create a factor Pool, with two levels corresponding to these groups, using **Edit>Factors>Add** (leaving entries for other sites blank), and run **Tools>Sum** on the counts sheet, giving factor Pool. On the resulting matrix, **Plots>Geometric Class Plot** clearly shows the right shift in the abundance distribution at the inner sites. Close the workspace; it is not needed again.



Dominance curves

Dominance plot is the convenient generic name for a family of curves also known as *ranked species abundance plots*, which can be computed for abundance, biomass, % cover or other biotic measure representing quantity of each taxon. For each sample, or pooled set of samples, species are ranked in decreasing order of (say) abundance. Their relative abundance (i.e. percentage of the total abundance in the sample) is plotted against the increasing rank (*x* axis), the latter on a log scale. The *y* axis can consist either of relative abundance or cumulative relative abundance, the former therefore always decreasing and the latter always increasing. The cumulative plot is often referred to as a *k*-dominance plot. There is a third possibility, a partial dominance curve, in which the *y* axis is the abundance of each species relative to the total of its own abundance plus that of all other less-abundant species. The idea of the latter is to ameliorate the way standard dominance curves tend to be dictated by the most abundant species, by looking at the dominance pattern of the remaining assemblage having removed the most abundant species, then the next most abundant, etc.

A further possibility is to put dominance curves for abundance and biomass, separately calculated, onto the same plot. This is referred to as an Abundance-Biomass Comparison (ABC) curve. A number of published studies have demonstrated a characteristic change in the relative position of these curves under disturbance, particularly for organic enrichment of marine macrobenthos, but a similar paradigm (loss of low-abundance large-bodied species and increased abundance of small-bodied ones) has been described for other faunal groups under impact (fish, birds, dragonflies, small mammals, etc). The method of display is due to Warwick RM 1986, *Mar Biol* 92: 557-562.

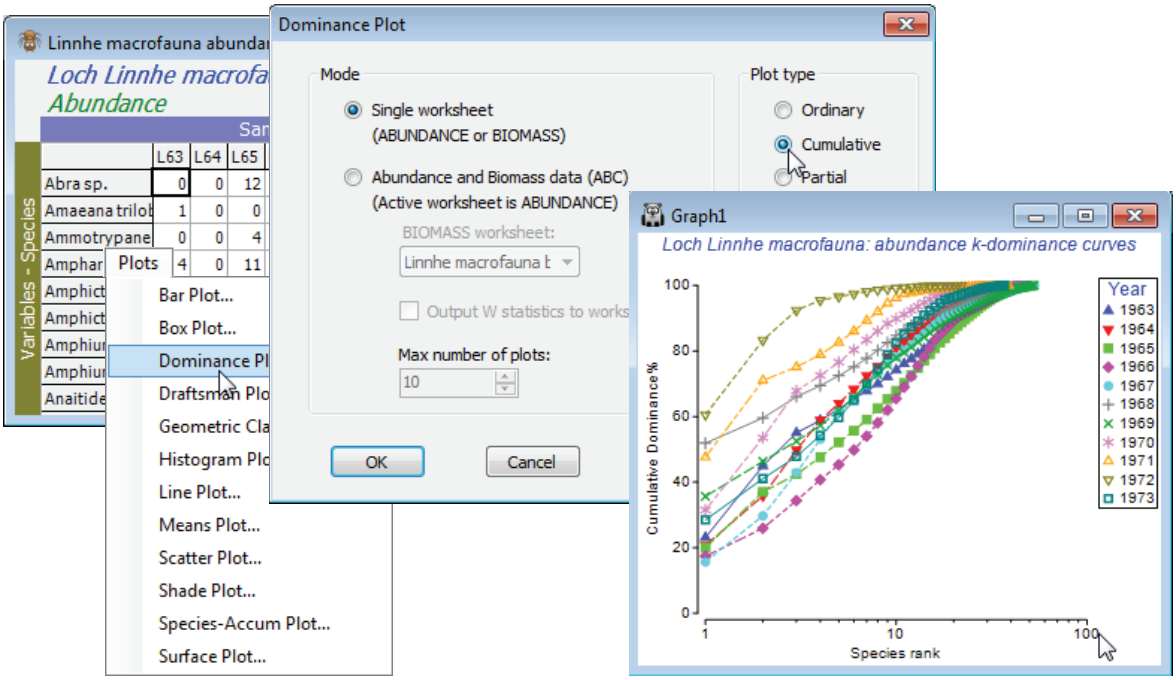
(L. Linnhe macrofauna time-series)

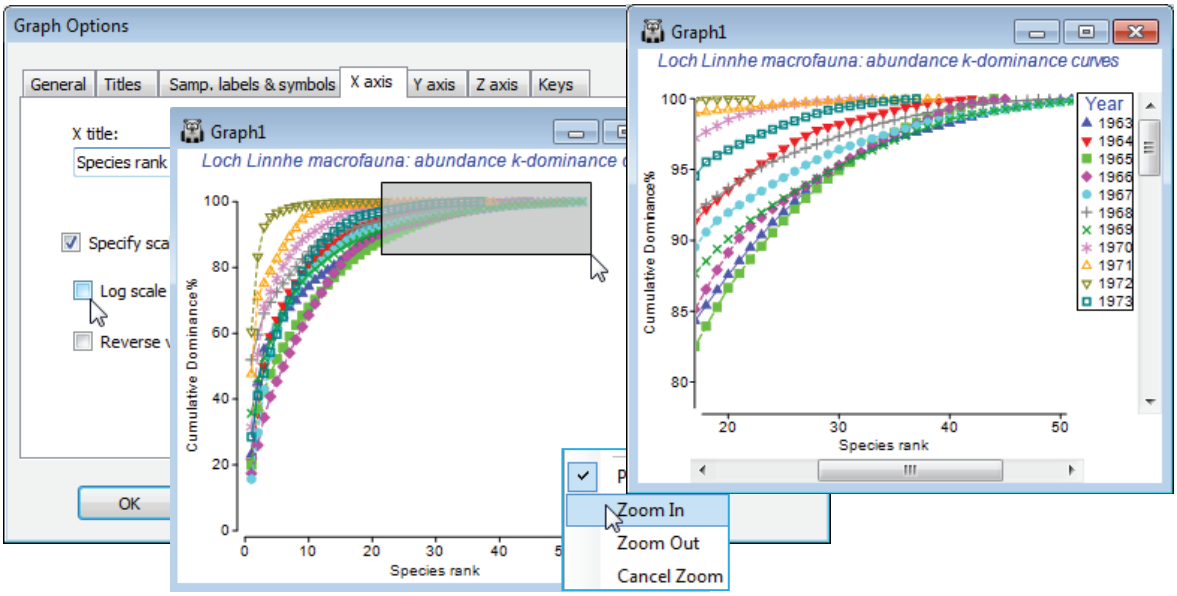
Macrobenthos in soft sediments of a site in Loch Linnhe, Scotland were monitored by Pearson TH 1975, *J Exp Mar Biol Ecol* 20:1-41, over the period 1963-73, recording both species abundance and biomass. The data are pooled to a single sample for each year, with an assemblage of 111 species. Starting in 1966, pulp-mill effluent was discharged in the vicinity of the site, increasing in 1970 and reducing in 1972. Abundances are in Linnhe macrofauna abundance, and the matching total biomass, of each species for each of 11 years, in Linnhe macrofauna biomass, in C:\Examples v7\ Linnhe macrofauna. Though Linnhe ws was created in Section 10, you may prefer a new start here.

k-dominance, ordinary & partial plots

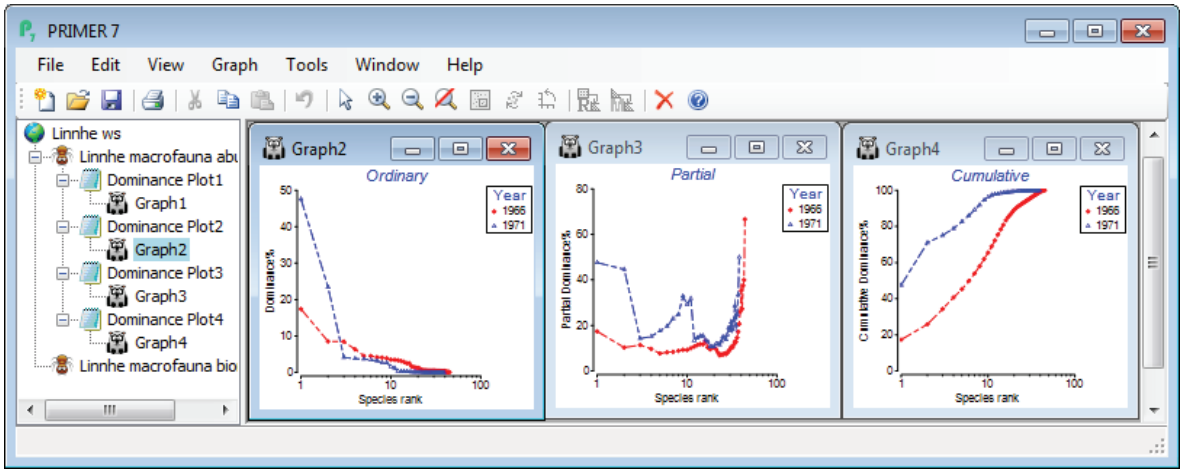
On the abundance matrix take **Plots>Dominance Plot>**(Mode•Single worksheet) & (Plot type• Cumulative). This produces a single plot of the *k*-dominance curves for all 11 years, the earlier years being seen to have higher evenness (lower dominance), by their placement lower down the *y* axis, and higher species richness by their extent along the *x* axis, before the full 100% is reached. In contrast, the years of highest effluent release are characterised by low evenness (curves move up the plot) and lower richness. The latter tends to be under-emphasised on *k*-dominance curves since the default is to plot species ranks on a log scale. This can be removed by unchecking (✓Log scale) on the **X axis** tab of Graph Options, best accessed simply by clicking the *x* axis scale, but typically that would underemphasise the equitability component, so a log scale is usually preferred. Note that most of the usual plot features, such as a rectangular zoom (changing the aspect ratio) are available.

v7 !





To note the effect of switching from cumulative to ordinary/partial dominance curves, first select just the years 1966 (the last year before effluent discharges began) and 1971 (one of the two peak years of discharge), by highlighting those columns in the data and **Select>Highlighted**. Perform three separate runs of **Plots>Dominance Plot**, with Plot type options of •Ordinary, •Cumulative and •Partial, then close all windows except the three graphs and **Window>Tile Vertical** to obtain the displayed desktop. Note how the cumulative plot emphasises the greater dominance of the two most abundant species in 1971 (once high % values are reached they cannot decrease!), but the partial plot also picks out an unevenness for that year in the dominance structure of species lower in the rank order, an undisturbed partial dominance curve typically looking more like that for 1966.



Abundance-Biomass Comparison curves

ABC curves plot abundance and biomass *k*-dominance lines on the same plot, and are interpreted in the literature as indicating an *undisturbed* community if the biomass curve is above the abundance curve, *gross disturbance* if the abundance curve lies above the biomass and *moderate disturbance* if the two largely intersect. This is based on the observation that for climax communities of soft-sediment macrobenthos the biomass dominants are large-bodied but do not dominate abundance, and are the more susceptible species to impact, whereas gross disturbance, especially from organic enrichment, leads to abundance dominance by a few, small-bodied opportunist species. (A very different example is given by Smith WH & Rissler LJ, 2010, *Restor Ecol* 18, 195-204, who show that ABC curves for herpetofauna track succession after forest fires).

Restore the full set of samples for Linnhe macrofauna abundance by **Select>All** (and **Edit>Clear Highlight** if you wish, though the latter is unnecessary since all routines – with the exception of **Pre-treatment>Transform (individual)** – operate on the current selection, not the highlights). The eleven ABC plots, one for each sample (year), are generated by a single run of **Analyse>Dominance Plot**, and placed in a multi-plot. The active sheet must be the abundance matrix and the Linnhe macrofauna biomass matrix must be available in the workspace as the secondary sheet.

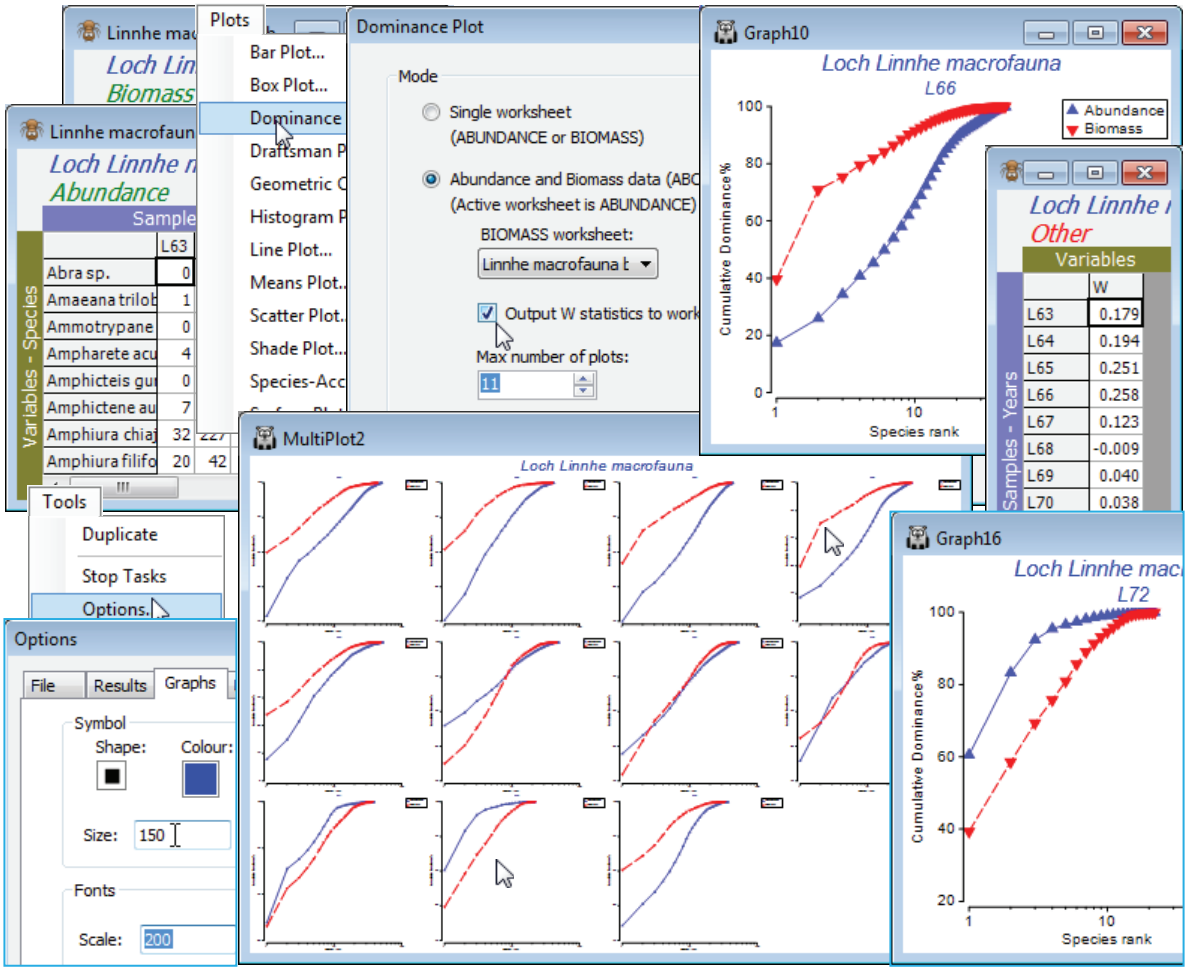
Matching when there are selections

An attempt to run the routine with the matrices reversed, or to run it on a data type *Environmental* sheet, will provoke one or more warnings (e.g. *Primary data not Abundance*), though not usually an outright error. It is always advisable to check that Data types have been correctly defined, and change them if necessary with **Edit>Properties** – you will then get the benefit of sensible defaults and warnings if you make an unexpected choice. As with all routines using a secondary sheet, there is strict matching of sample labels in combining the sheets, but with the option to relax this and take samples in the worksheet orders, if the two matrices are of the same size. If they are not, and all the abundance sample labels can be found in the biomass sheet, then plots will be performed for all the samples with abundance data. (Thus, if the selection of years 1996 and 1971 in the previous illustration had not been rescinded, the routine would return ABC plots only for those two years).

v7 This is in line with the general principle throughout v7 – the active sheet determines which samples are analysed, the secondary sheet is a look-up table (but it is advisable in this context to make sure everything matches!). The species labels do need to match because a careful check is made that there are, for example, no species which have a positive biomass but zero abundance (the converse is permitted, since there may be species which are present but too small-bodied to have measurable weight). After those checks are completed satisfactorily however, the dominance curves re-order the species differently (in decreasing rank order) for each sample and for biomass and abundance – it is inherent to the method that the species are not in the same rank order for the B and A curves.

So, on the abundances, take **Plots>Dominance Plot>**(Mode•Abundance and Biomass data, ABC)> (Biomass worksheet: Linnhe macrofauna biomass) & (✓Output W statistics to worksheet) & (Max

v7 number of plots: 11) & (Plot type•Cumulative). In addition to the multi-plot of 11 ABC plots there will be a column of *W* statistics for each sample. *W* measures the extent to which the biomass curve lies above the abundance curve (positive for *undisturbed*, negative for *impacted* samples, in theory) and is a convenient single index to report, if presenting large numbers of ABC plots is impossible. Note also in the below that the font and symbol sizes in the single plots (obtained by clicking on that component in the multi-plot) are enlarged. This is much more conveniently done in a single operation, prior to running **Dominance Plot**, by changing defaults with **Tools>Options>Graphs**. These new defaults are in action for all plots (until changed), even after exit from PRIMER.



The *W* sheet can then be **Tools>Merge**(d) with other classic or distinctness-based diversity indices for univariate plotting, testing etc. Sometimes the juxtaposition of abundance and biomass data in *W* does capture a different diversity dimension than the classic axis of richness-evenness seen for the multivariate analysis of diversity measures in the previous section. The ABC plots in this case follow the pattern, seen in other diversity indices also, of initial stability (with the biomass curve over abundance), then a gradual switch over of the curves in the period of effluent impacts from 1966, increasing in 1970 (abundance clearly over biomass in 1972), with apparent recovery in 1973 after the discharges decreased in 1972. Close the Linnhe workspace; it will not be needed again.

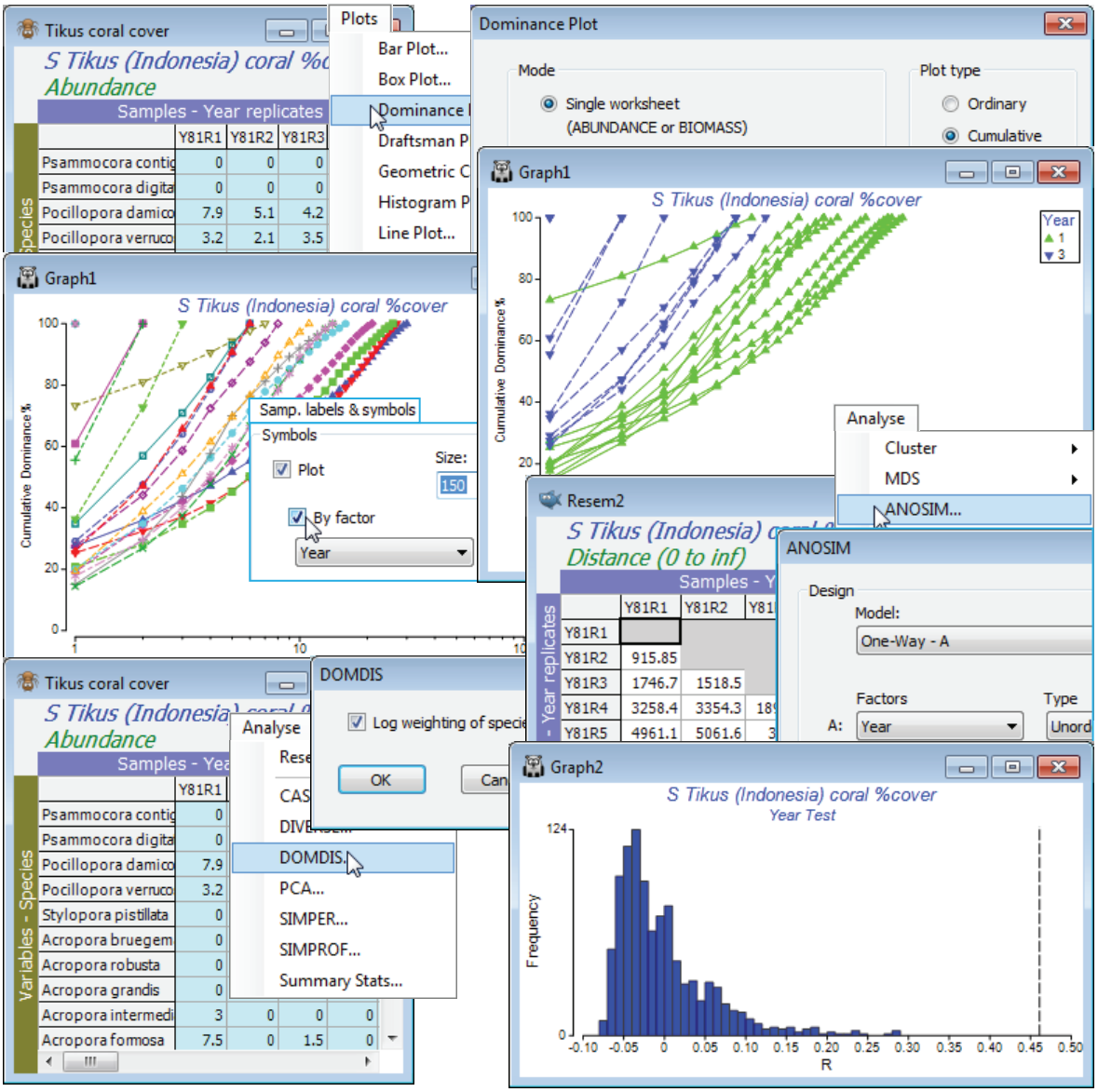
Testing for *k*-dominance curves

Testing for differences in ABC curves for group structures of sites, times or treatments etc, where there are replicate samples within each group, is probably best accomplished by using the *W* index. This is computed for every replicate and the *W* values are treated like any other diversity measure, by univariate statistics. A different approach is needed for *k*-dominance curves, because of the lack of an internal comparison of curves to generate a univariate statistic. Single cumulative curves now need to be compared across replicates, both within a group and between groups. Clarke KR 1990, *J Exp Mar Biol Ecol* 138: 143-157 suggests a solution here, which is implemented in the **Analyse> DOMDIS** routine. This starts from an active sheet of a single species×samples array (abundances, for example, though it could equally be biomass or area cover as in the example that follows), then calculates separately for each sample the cumulative relative abundances of species ranked in their decreasing order, as for the *k*-dominance plots. The distances apart of all pairs of cumulative curves (samples) is now computed, using Manhattan distance D_7 (see Section 5), and the routine therefore generates a resemblance matrix (dissimilarity of curves) among all samples. This can be entered into the multivariate PRIMER routines in just the same way as for any other dissimilarity matrix. (The possibility of inputting pairwise distances between curves – growth curves, PSA curves etc – to multivariate analysis was seen in Sections 4 and 5). In particular, a run of **Analyse>ANOSIM** on this distance matrix will produce a significance test for the differences among groups. Replicate curves across groups that tend to be further apart from each other than replicates within groups will give ANOSIM $R > 0$, and this is tested by permutation as usual. In fact, there is a choice of two *curve separation* statistics offered by the dialog box in **DOMDIS**, namely Manhattan distance and a modification of it, which is the default: (✓Log weighting of species ranks). This multiplies the absolute difference between the curves at the *i*th point on the x axis (the *i*th ranked species) by $\log(1 + i^{-1})$, which successively downweights the contributions from the lower ranked species. It reflects the fact that *k*-dominance curves are usually plotted with a log scale on the *x* axis (of ranks), and it approximates to the visually-observed area between the two curves. The unweighted form would be relevant if plots are used without a logged *x* scale (as seen in an earlier example).

(Tikus Is coral cover)

The Tikus Island, Indonesia, data on % area cover of coral communities on 10 transects in the years 1981, 83, 84, 85, 87 and 88, were first met in Section 5, with data sheet Tikus coral cover in the directory C:\Examples v7\Tikus corals. It is discussed extensively in the context of resemblance choice for multivariate analysis in Chapter 16 of CiMC but here, as an illustration of *k*-dominance curve testing, we will use only the differences between two of the years: 1981 and 1983, before and after the major El Niño event of 1982/3.

Open Tikus coral cover in a new workspace and **Select>Samples>**(•Factor levels)>(Factor name: Year)>**Levels** and Include only 1 and 3. **Plots>Dominance Plot**, with the defaults of (Mode•Single worksheet) and (Plot type•Cumulative), will generate the *k*-dominance curves for all the selected samples on a single plot. Show the group structure of 10 replicates from each of two years by **Graph>Sample Labels & Symbols**, switching on (Symbols:✓Plot>✓By factor: Year). To test for significance of the (rather obvious!) differences between the curves in the two years, run **Analyse> DOMDIS>**(✓Log weighting of species ranks) on Tikus coral cover. This generates a resemblance matrix Resem1, giving the distance between all pairs of curves, which is then input to **Analyse> ANOSIM>**(Model: One way - A)>(Factors A: Year)>(Type: Unordered), with the other defaults taken. The *R* statistic is 0.46, easily larger that for any of the 999 random permutations under the hypothesis of no year difference, thus p<0.1% (and the null histogram shows it would be a great deal smaller, for more simulations), signifying a clear change in dominance pattern between the years. The *R* value is depressed somewhat by one outlying replicate in 1981 which is much more dominated, and less species rich, than the other 9 transects, but there is no justification for leaving it out of the analysis. Close the workspace – it will not be needed again.



(Sea-loch contiguous macrofauna cores)

The final set of data in this section is of a benthic study by Gage JD & Coghill GG 1977, in Coull B (ed) *Ecology of marine benthos*, Univ S Carolina Press, at a single site (C-12) in Loch Creran, Scotland, involving 256 contiguous cores arranged along a single transect. Small cores were used to examine local-scale dispersion (clumping) properties of sediment macrofauna. The data matrix of 67 species by 256 samples is the file Creran macrofauna counts in directory C:\Examples v7\Sea-loch macrofauna; open this into a new workspace. It will be used here to illustrate the final type of curvilinear plots available in PRIMER, *species accumulation curves*.

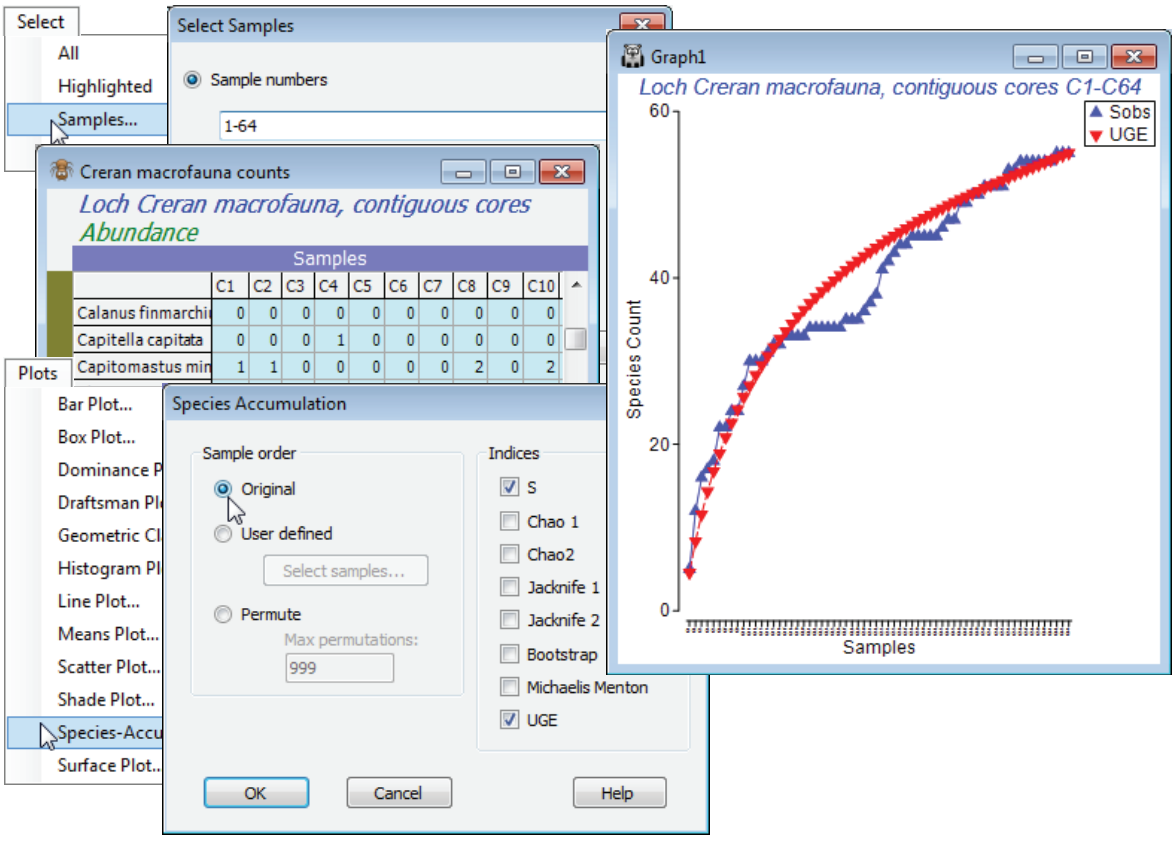
Species accumulation plots

The **Plots>Species Accum Plot** routine plots (and lists) the increasing total number of different species observed (*S*), as samples are successively pooled (often referred to as the *S obs* curve). This is accessed when the active sheet is any species × samples matrix (though for all except the Chao1 index, see later, which requires genuine counts, only the presence/absence structure is used). There are three options for the order in which samples in the data sheet are successively amalgamated: (Sample order•Original), (•User defined) or (•Permute). The first case simply takes samples in their label order in the worksheet and the second specifies the order by a **Select samples** button. This provides an Ordered Selection dialog, in which samples can be moved up or down with the usual ↑ ↓ buttons – alternatively, re-order the original sample labels with **Edit>Sort**, using a factor. The third (default) option is to enter samples in random order, this being carried out 999 times (or whatever specified) and the resulting curves averaged, giving a smoothed *S* curve.

The analytical form of this mean value of the accumulation curve (over all permutations, in effect) was given by Ugland K, Gray JS, Ellingsen K 2003, *J Anim Ecol* 72: 888-897, and is computed by

Plots>Species-Accum Plot>(Indices✓UGE). [It is the counterpart of the analytical form given by Hurlbert SH 1971, *Ecology* 52: 577-586, for the Sanders rarefaction curve met under **Analyse> DIVERSE**, which gave expected numbers of species for subsets of *individuals* from a single sample (whereas here we are talking about subsets of *samples* from a data matrix).] Although, for large numbers of permutations, the (✓UGE) curve will always lie on top of that from specifying (Sample order•Permute) & (Indices✓S), it is included as a separate option so that the combination of original sample order for *S* with the mean curve (UGE) can be taken. For samples that arrive in a non-arbitrary space or time order, this allows comparison of the real accumulation curve with its smoothed version – spatio-temporal heterogeneity will display as a jagged or stepped *S* curve.

Select the first 64 samples from Creran macrofauna counts, either using **Select>Samples>**(•Sample numbers: 1-64) or (•Factor levels)>(Factor name: sixtyfours)>**Levels** and Include only level 1, and submit to **Plots>Species-Accum Plot>**(Sample order•Original) & (Indices✓S✓UGE), unchecking the other indices. There is perhaps a suggestion of stepping in the *S obs* curve (though not much, and this would be hard to test for). However, it is clear that the accumulation curve is still rising and this raises the question as to how much larger *S* can get, with repeated sampling in this area.

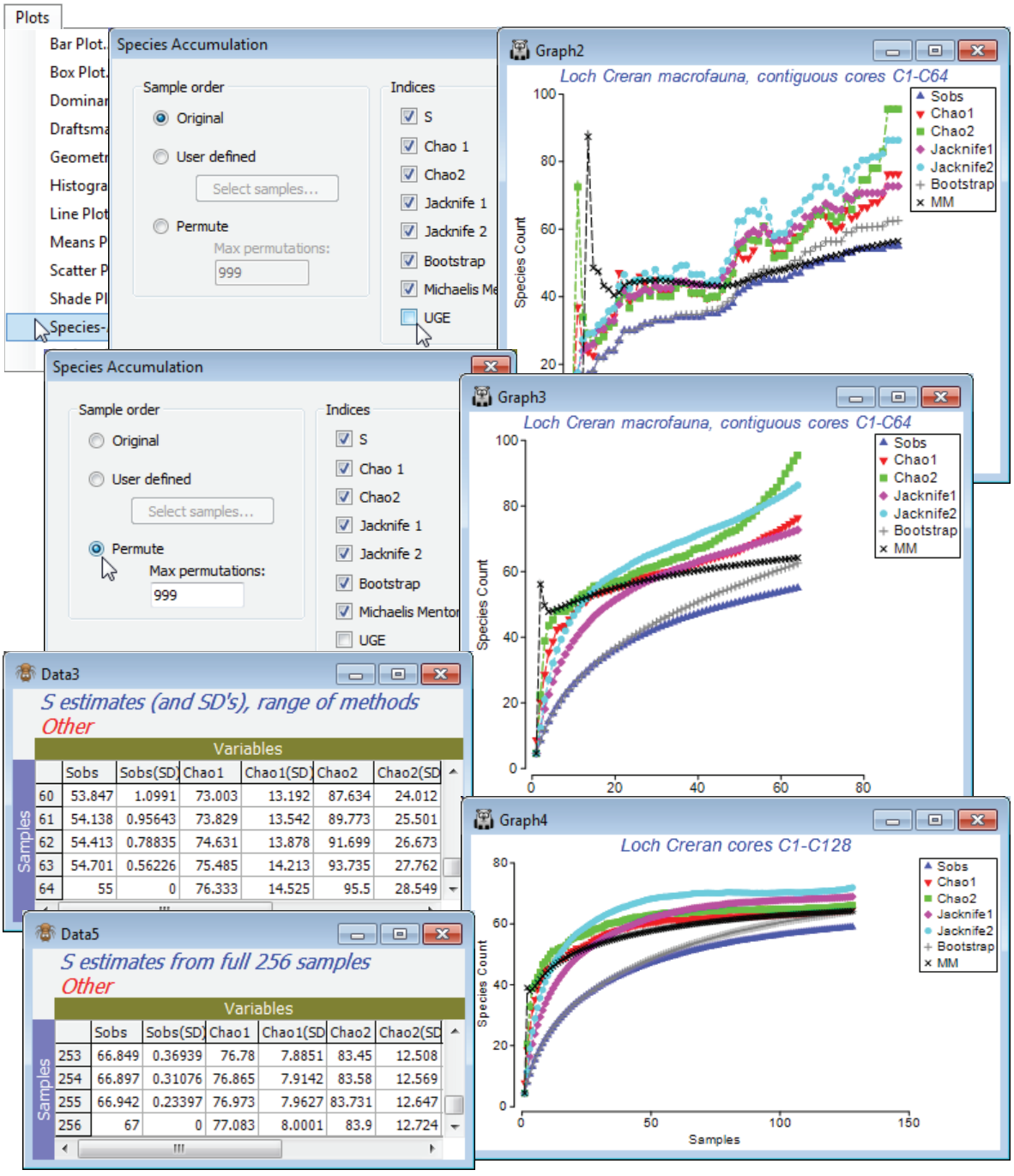


S estimators

PRIMER therefore includes a number of *S extrapolators* – attempts to predict the true total number of species that would be observed as the number of samples tends to infinity (the *asymptote* of the species accumulation curve), assuming that a closed community is being successively sampled. This should not be confused with the (✓UGE) or permuted (✓S) curves which, like rarefaction indices, look backwards at the expected behaviour of *S* as samples are removed, and return simply the observed *S* at the end of the series. There is a choice of six extrapolators, each of which is calculated as every new sample is added, so the result is again a curve, of the evolution of the *S* predictor as sample size increases (though mostly one would use the end point prediction as the best estimate of the asymptote). Where the samples are entered in permuted orders, the predictions are again the average of the 999 estimators at each step. These are not parametric approaches, depending on simple functions of the number of species seen only in 1 or 2 samples (Chao2, Jacknife1 and 2), or the number of species that have only 1 or 2 individuals in the entire pool of samples (Chao1), or the set of proportions of samples that contain each species (Bootstrap). The *v7* only parametric model given is Michaelis-Menton, which returns (more appropriately, in PRIMER 7!) the predicted asymptote of a hyperbola fitted to the cumulative *S* curve at each step.

The literature on *S* estimation is large, and PRIMER does not attempt a comprehensive approach. An early and influential summary for ecologists is Colwell RK & Coddington JA 1994, *Phil Trans Roy Soc B* 345: 101-118, who detail the above estimators, and an excellent software package for serious users in this area is Colwell's 'EstimateS' (http://viceroy.eeb.uconn.edu/estimates/).

So, produce these six differing *S* estimates with **Plots>Species Accum Plot** from samples 1-64 in the Creran macrofauna counts sheet, in two ways: with the samples in their original order and run again with (Sample order•Permute)>(Max permutations: 999), unchecking (✓UGE). In addition to the plots, there is a worksheet of the numeric estimates at each step, which includes (in some cases) simple standard errors based on the permutations – but some of these are certain to underestimate the true degree of uncertainty in practice. (Of course, determining the true number of species not seen is an essentially unsolvable problem without strong assumptions about closed communities and catchability by the sampling device, which in a marine context can be doubtful – accumulation curves rarely approach asymptotes quickly). Check the estimates made at 64 samples against the number of species observed in 128 samples and, run again at that level, against the total *S* for the full set of 256 contiguous cores. Has an asymptote effectively been reached? Close the workspace.



17. Bootstrap regions for group means (*Bootstrap Averages)*

Analogue of univariate means plots

For this final section, we consider only cases in which the samples form a 1-way layout, i.e. there is a single (*a priori* defined) factor whose levels divide the data into groups, the samples in a group being considered replicates of that factor level. This factor could, of course, be from a combined factor for a higher-way crossed layout, e.g. all combinations of sites and times with (real) replicates for each combination. For a univariate response, such as a diversity index, the first step would be to test if there were significant differences among the groups and, if so, this is followed by a *means plot*, showing the respective means for each group with some measure of how reliably that mean has been determined, usually a confidence interval surrounding the mean. This is the type of output generated by all statistical packages (and by the **Plots>Means Plot** routine, seen in Section 15 for the taxonomic distinctness diversity index). It is uncommon, for example, to see a plot of the actual replicate values juxtaposed for each group in the univariate case. A glance at Fig. 18.1 of CiMC will soon tell you why! By comparison with a means plot, displaying the replicates will often make it difficult to see that pattern among the means clearly. Yet, in the multivariate context, there seems to be a (false) perception that the correct ordinations to plot have always to be ones that contain a point for each sample, in spite of the high stress solutions that often result – with that stress often due to random sampling variability within each group rather than to an inherently high-d structure in the relationship among groups. *n*MDS is surprisingly successful in capturing underlying group structure from replicate data, partly because it is able to non-linearly stretch and squeeze the dissimilarities into the low-d MDS distance scale, compressing the scale for smaller dissimilarities, which are often those from within groups. However, the logic from univariate practice (and it is sound logic) is that for *a priori* defined structures, we should run a testing procedure (ANOSIM, RELATE, PERMANOVA, …) to justify interpreting the data and follow this with an ordination means plot, to display and interpret the among-group patterns. On several occasions in this manual and in CiMC, we have followed through this logic to good effect. We have also seen that there are a number of ways of producing a means plot, i.e. different ways of defining a *measure of central tendency* (to use the accepted jargon of statistics). This is particularly so in the multivariate context, where we have already referred to relationships among means calculated from: averaging the raw data, the transformed data matrix, the similarities, the rank similarities, and (in PERMANOVA+) generating centroids in the high-d PCO space. Chapter 18 of CiMC discusses the pros and cons of some of the different methods, but the practical reality (fortunately) is that the relationship among means is often very similar whichever method is used. The possible exception to this is averaging raw data in cases where a strong transformation is appropriate – the means can then get hijacked by species with large outlying counts in a single replicate, something which the usual process of transformation before computation of similarities is able to resolve. Ordination plots of means, based on averaging a data matrix with **Tools>Average** (rather than averaging biological-type similarities – also by **Tools>Average**, but on the resemblance matrix) are therefore preferably better carried out on the transformed form of that data sheet. (This is especially true if that averaging is over the levels of other factors.)

v7

v7

Status of region estimates

Ordination means plots are therefore a vital tool for interpretation but, in relation to their univariate counterparts, they lack one useful feature – the ability to get an approximate feel for the uncertainty in our knowledge of position of each mean point on that plot. In the univariate case, this is provided by 95% confidence intervals, or error bars of ± 1 or 2 standard errors (i.e. standard deviations of the mean). These do not reflect primarily the variability in individual replicates from a group but (in a way that is rather precisely defined but rather loosely understood!) the uncertainty in knowledge of the true mean, and this is largely dictated by the number of replicates. Of course, such interval estimates, and the confidence values, are based on parametric assumptions about the distributional form of the observations. This can be carried over into, for example, a concept of 95% confidence *regions* within an (*x,y*) plot of groups for a two-variable matrix, but increasingly inflexible and unrealistic assumptions need to be made (the standard multivariate normal MANOVA model). It is realistic, for biological assemblage matrices typically in high-d space and with complex and largely non-identifiable parametric dependencies among the species variables, to set our sights lower and seek to display approximate regions around each group mean, e.g. on a 2-d ordination, which give a 'feel' for the uncertainty in that mean's position. The regions do not have the status of confidence intervals therefore, and should not be used as tests – that is the role of ANOSIM etc.

Bootstrap definition

The construction of approximate regions is approached through *bootstrap averages*. We only have one observed mean from *n* replicates for a particular group – leaving aside how that is calculated for the moment. What we need, for a plausible region estimate, is examples of what 'other means might have looked like' had we been able to repeat exactly the same sampling protocol, e.g. to obtain another set of *n* replicates at the same time and/or place (or whatever the group represents). We also need the 'other means' to be generated without distributional assumptions, in keeping with the rest of our approach. There is no possibility of obtaining this by permutation (permuting the sample labels across groups destroys the group structure we are trying to represent, and permuting within a group does not change its mean!) However, resampling the set of *n* samples *n* times, with replacement, will produce a different sample set (some samples are certain to be missed altogether and some are repeated, possibly a few times), and will have a different mean. This is a *bootstrap* sample, giving a single *bootstrap average*, and repeating this process *b* times will give a set of *b* such averages which are, to some degree, a set of 'other means which we might have obtained'.

v7

Bootstrap regions

These averages can then be used to generate a *bootstrap region* for each of the *g* groups – at its simplest by displaying the full set of *b*×*g* averages in a 2-d (or 3-d) ordination. Here is the first obvious approximation therefore, namely that a 2- or 3-d ordination is not necessarily a perfect representation of the *b*×*g* samples, since they are from a higher-d variable space. But this is an issue we are well used to dealing with – we interpret 2-d ordinations cautiously if they have high stress, and look at the 3-d plots (or even subsets of higher axes, though this is rarely necessary in this case) to check whether the 2-d plot has over-simplified some aspect of the groups' structure. In fact, the stress values for a 2-d plot are often quite acceptably low, even though these are typically ordinations on a very large number of samples of bootstrap averages (the recommendation is *b*=100+ bootstraps per group, if you can run this in a viable time, i.e. an ordination on 500+ points if you have *g*=5 groups). This is because the inherent structure of the plot may be just that of the relationships among the *g* group means, and such means plots are usually low-dimensional. At least this will be the case if the original number of replicates per group is not small, so that the regions are fairly tight (and PRIMER will issue a warning if you run **Analyse>Bootstrap Averages** with groups which are definitely too small – less than 5 replicates, though many more are preferable).

v7

v7

The **Bootstrap Averages** routine is able to take this a stage further and, for the 2-d ordination, will construct smooth envelopes for the bootstrap average points which have a nominal 95% coverage (or 80% or 50%). As stressed above, this is not a formal 95% confidence interval, since several sources of uncertainty (such as the approximation to the 'true' dimensionality) are not catered for, but a subtle and rather complex correction is made for the well-known underestimation (of order $1\text{-}n^{-1}$ on both axes, where *n* is the number of original replicates in a group) in variance estimates from bootstrap means. The nominal 95% coverage comes from approximating the shape of the observed bootstrap average regions in 2-d by back-transformed bivariate normals from individual location-shifted power transformations, fitted to the rotated major and minor axes for each group separately (essentially the algorithm used in Section 17 for Δ^+/Λ^+ 'ellipse' plots). This is another approximation therefore, and will not be able to fit non-convex (e.g. banana-shaped) clusters of points very convincingly – but it does incorporate the variance bias correction, so it is generally seen that the smooth envelopes contain more than 95% of the bootstrap average points.

Metric or non-metric plots?

v7

The **Analyse>Bootstrap Averages** routine allows both metric and non-metric options for the MDS ordination of the bootstrap average regions. However, *m*MDS is the recommended choice, and the default. The motivation for constructing a region plot for the group means, rather than just the point estimates of a simple means plot, is to allow interpretation of the among-group structure in relation to the uncertainty in the positions of group means (exactly as we use interval estimates in univariate studies). It is useful to be able to visualise that, for example, along a line connecting two group means A and B, the degree of uncertainty in mean A is about 20 dissimilarity units, in mean B it is about 10 (perhaps B has more replicates or smaller innate dispersion – no assumption of common dispersion or balanced designs need to be made, see later) but at their closest point the two regions are still 20 dissimilarity units apart. This requires the linear measurement scale of a metric MDS (*m*MDS, not *n*MDS or even *tm*MDS, see Section 8). *m*MDS solutions at the level of replicate samples can often be very poor, with high stress and representation of the among-group structures compromised by the need to display sampling error in full – this is why we use the much more flexible *n*MDS. But at the level of group means and their greatly reduced sampling variability by averaging, *m*MDS is often of acceptable stress (even with many bootstraps) and very interpretable.

Bootstrap averages in a reduced *m*MDS space

v7

Though hopefully the above gives the motivation and an idea of the way the region estimates are constructed, the most important instruction of this Section is to read Chapter 18 of CiMC! The detailed reasoning and information it gives will not be repeated here but the upshot is that the best way of constructing the bootstrap averages which are then displayed (and smoothed, bias-corrected etc) in 2-d *m*MDS, is to calculate them from m-dimensional metric MDS ordination co-ordinates created by running *m*MDS on the original dissimilarity matrix. This is carried out for a range of increasing values of m, starting from m=4 (up to m=10, though this limit is usually not reached) and stopping when the ordination configuration crosses a threshold for how well it matches the original dissimilarity matrix. In other words, m is chosen to be just high enough to give a 'near-perfect' representation of the dissimilarities. The criterion, as used in several guises in previous sections (the cophenetic correlation of cluster analysis, the matching coefficient for RELATE and BEST/BVStep, e.g. of a subset of species to the multivariate sample pattern for the full species set etc.) is just a matrix correlation ρ. Here this is between the original dissimilarities and the distances (which are Euclidean of course) among the sample points on the *m*MDS – in other words, ρ is the Pearson correlation of the points in the Shepard diagram. (In the context of *m*MDS and the need to retain the metric information in the original dissimilarities, as discussed above, it makes sense to use a standard Pearson correlation here and not the usual rank-based Spearman correlation). The default in **Analyse>Bootstrap Averages** for (•Auto m) choice is that the smallest m is chosen to make Pearson $\rho \geq 0.99$, though this is under user control. The threshold criterion we adopted for successful reconstructions of the original dissimilarities, in the BVStep runs at the end of Section 14, was (Spearman) $\rho \geq 0.95$, so the more severe $\rho \geq 0.99$ could certainly be relaxed a little if necessary, without compromising the approach. As shown in Chapter 18, CiMC, the dimension m in which the bootstrapping operates must avoid being too large, otherwise an artefact of high-d bootstrapping becomes increasingly important, resulting in significant underestimation of true dispersion by the bootstrap averages, however many original replicates there are in a group (i.e. however well-behaved a univariate bootstrap might be). This explains the restriction to $4 \leq m \leq 10$ in the (•Auto m) option, but the routine also permits manual choice of m, to allow the user to look at the outcome from a wider range of dimensionalities.

v7

Starting from an active sheet which is the full sample resemblance matrix, **Analyse>Bootstrap Averages** therefore replaces this by m-dimensional *m*MDS co-ordinates (another approximation therefore in the series leading to our smoothed, nominal 95% region estimates! – but a very useful one, giving the technique some excellent properties). It is in this reduced space that *n* bootstrap samples are chosen, for a group with *n* replicates (*n* will differ for each group, in general) and their means calculated – so the *Bootstrap Averages* of this section are all simple averages for each of the m co-ordinates of an *m*MDS ordination. This is repeated *b* times – also under user choice, though the routine suggests a default which limits the overall number of bootstrap averages across all groups to 300. (However, most machines can run MDS for at least twice that number, hence the earlier encouragement to increase *b* to at least 100, if at all feasible).

Output options for region plots

v7

These *b* bootstrap averages for each of the *g* groups are then displayed in a low-d *m*MDS space and you have control of whether to display any or all of: the *b*×*g* bootstraps (✓Bootstrap averages); the overall averages for each group (✓Group averages); the smoothed envelopes (✓Bootstrap region (2D only)). You cannot easily change your mind about the choice of what is on the display after the run has completed so you may find yourself running the routine more than once, with differing display options. The (✓Group averages) are not the centre of a group's points in the (usually) 2-d *m*MDS space of the final display, but the centre of gravity in the m-dimensional space in which the bootstrap averages were calculated, and which are then ordinated into 2-d along with the bootstrap averages. The theoretical unbiasedness of bootstrap averages ensures that this is essentially the same as the centre of gravity of the original samples for that group, after they have been placed in the m-dimensional *m*MDS space. These group average points can only therefore fail to lie in the centre of the displayed bootstrap averages if there is some distortion in going down from the m-dimensional space of the calculations to the 2-d (say) space of the display, which is potentially useful information (and might suggest looking at the structure in 3-d). This also highlights another important consequence to carrying out the bootstrapping in the reduced m-dimensional *m*MDS space: because these are simple averages of points in Euclidean space, unequal replication across the groups is not a problem – it will not give the bias that bootstrap averaging in the species space is likely to face (i.e. larger samples produce more species and this changes similarity structures).

(W Australia fish diets)

The diet study for 7 species of W Australian fish, with a variable number of dietary samples from each, has been seen several times now. In Section 3 we selected a subset of just the 3 congeneric species, *Sillago schomburgkii* (10 samples), *S. bassensis* (14) and *S. vittata* (16), labelled as B, E and G, and decided to omit samples B3 and B4 from these species (and also A9 from *A. ogilbyi*) on grounds of very much lower total gut content than other samples. In Section 4, we noted that it was necessary to standardise these samples across the 32 prey categories (the 'species' variables) since total gut content of a sample (units are %volume) could not be controlled – since the fish are doing the sampling(!) – and differences in these totals are of no relevance in seeking dietary distinctions among these 3 closely-related species. In Section 8, we looked at *n*MDS plots for all 7 fish species in higher-dimensions since a 2-d plot had high stress, and noted the differences in variability in diet among the species, and in section 9, we ran the ANOSIM global and pairwise tests between the species to establish the statistical significance (or otherwise) of dietary differences. In Section 10 we referred to use of SIMPER to identify the main dietary categories accounting for dissimilarity among fish species, where these were established with confidence by the ANOSIM tests (and shade plots would also be instructive here). Previously, towards the end of Section 8, we had shown some of those dietary categories on the 3-d *n*MDS of the samples, in a 3-d bubble plot, and also gave a means plot, averaging over the samples for each of the 7 fish species – on this we then displayed a segmented bubble plot of the main prey categories distinguishing the diets. We can now fit one final small piece to this jigsaw, by adding bootstrap regions to the means plot. In fact, it would be unwise to attempt this for the full set of species, since three of them have only 3, 4 and 5 replicate samples and the motivating concept of bootstrapping (generating 'other sets of means we could have observed') is more or less certain to break down with such small replicate numbers – the sampling with replacement just does not produce enough realistically different combinations, see Chapter 18 of CiMC, to avoid major underestimation of the true variation in those means. A fourth species, *A. ogilbyi* does have plenty of replicates and could be included (and you could do so), but for this illustration we look again at a logical subset – the three congeneric *Sillago* species – and construct an *m*MDS plot with bootstrap regions for their average dietary assemblages.

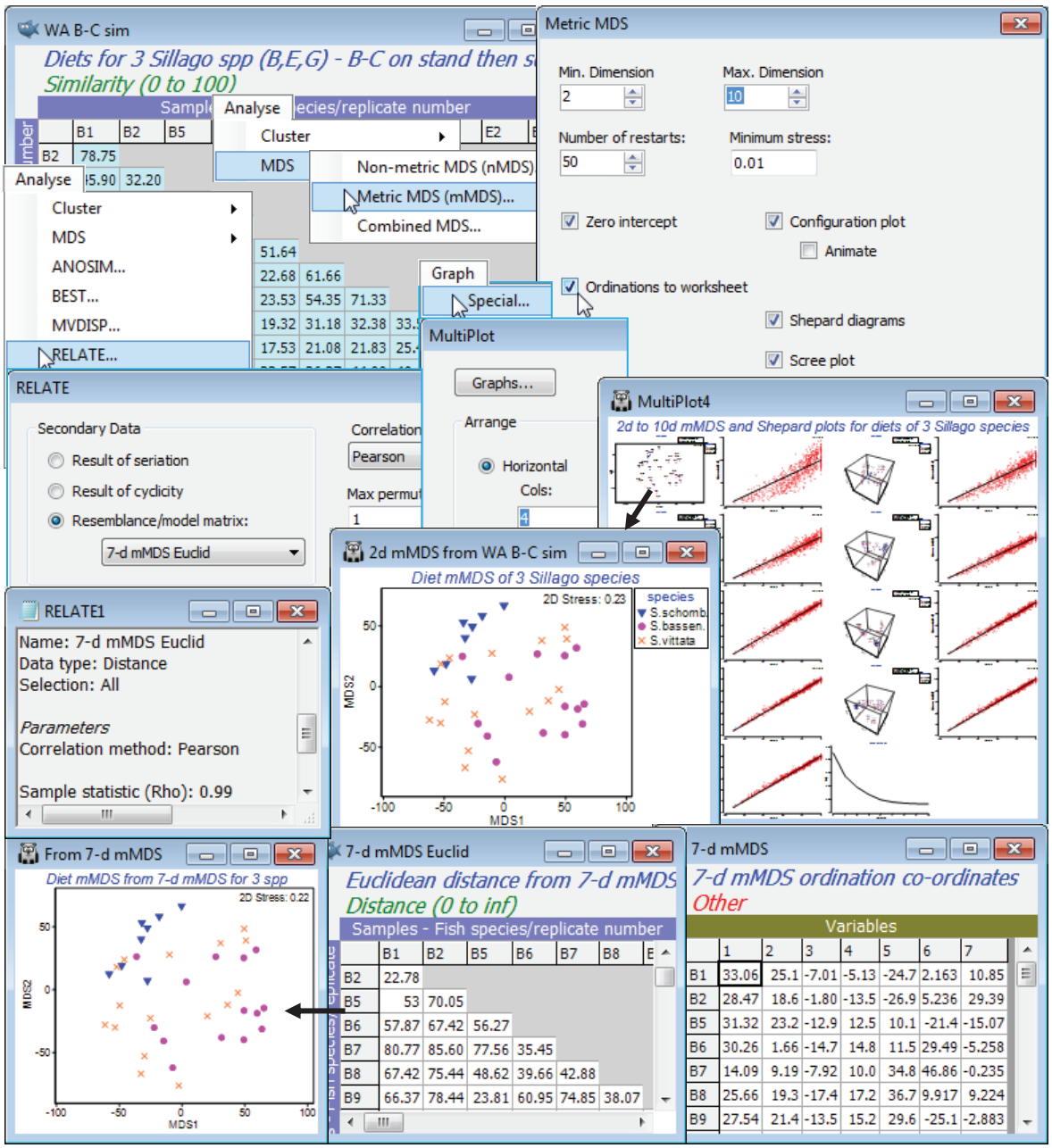
v7

v7

Open workspace WA fish ws from C:\Examples v7\WA fish diets, which should contain the resemblance matrix WA B-C sim for all samples (excluding A9, B3 and B4) of the 7 fish species. If not available, you will have to repeat the earlier steps of: opening the data file WA fish diets %vol, deselecting A9, B3, B4, standardising samples by total (**Pre-treatment>Standardise**), square root transforming and computing Bray-Curtis, renaming this WA B-C sim. Now highlight and select from this just the B, E and G labels – or instead **Select>Samples>**(•Factor levels)>(Factor name: species)>**Levels**>(Include: S.schomb., S.bassen., S.vittata). On these selected resemblances, run **Analyse>MDS>Metric MDS(mMDS)>**(Min. dimension 2) & (Max. dimension 10), taking other defaults such as (✓Zero intercept), but adding (✓Scree plot) and (✓Ordinations to worksheet). On the multi-plot this produces, reform the rows/columns by **Graph>Special>**(Arrange•Horizontal) >(Cols: 4). By clicking on individual plots within this multi-plot, note how the 2-d *m*MDS is of relatively high stress for the (only) 38 samples – as noted earlier, *m*MDS stress values will always be higher than for the equivalent *n*MDS but the Shepard diagram for the 2-d plot is also not very convincing. This changes, however, for higher dimensional solutions, with the Shepard diagrams becoming increasingly well described by a straight line through the origin (read across rows of the multi-plot for the Shepard diagrams in dimensions: 2 & 3, 4 & 5, 6 & 7, 8 & 9, 10 & the scree plot). By the time this gets as far as about a 7-dimensional *m*MDS solution, the stress has reduced to 0.03-0.04 – a very low stress for a metric MDS plot, capturing the original dissimilarities (not just their rank orders) to a high precision in the (Euclidean) distances between co-ordinate points in the 7-d *m*MDS plots. In fact, it is worth seeing that if you repeat the above 2-d *m*MDS, but this time starting from those Euclidean distances in 7-d *m*MDS space, the 2-d ordination appears to be identical to that produced from the original Bray-Curtis similarities for these three species. [You obtain these Euclidean distances by finding the data sheet containing the 7-d co-ordinates from your original *m*MDS run – because the (✓Ordinations to worksheet) instruction has sent 9 further sheets to the Explorer tree, of the 10-d down to the 2-d *m*MDS co-ordinates – then simply enter that 7-d co-ordinate data to **Analyse>Resemblance>**(•Euclidean distance), giving 7-d mMDS Euclid]. Of more relevance to understanding the bootstrap methodology, however, is to calculate (not test) the **Analyse>RELATE** statistic on the original similarities, WA B-C sim, with (•Resemblance/ model matrix: 7-d mMDS Euclid) & (Correlation method: Pearson), which returns $\rho = 0.99$.

v7

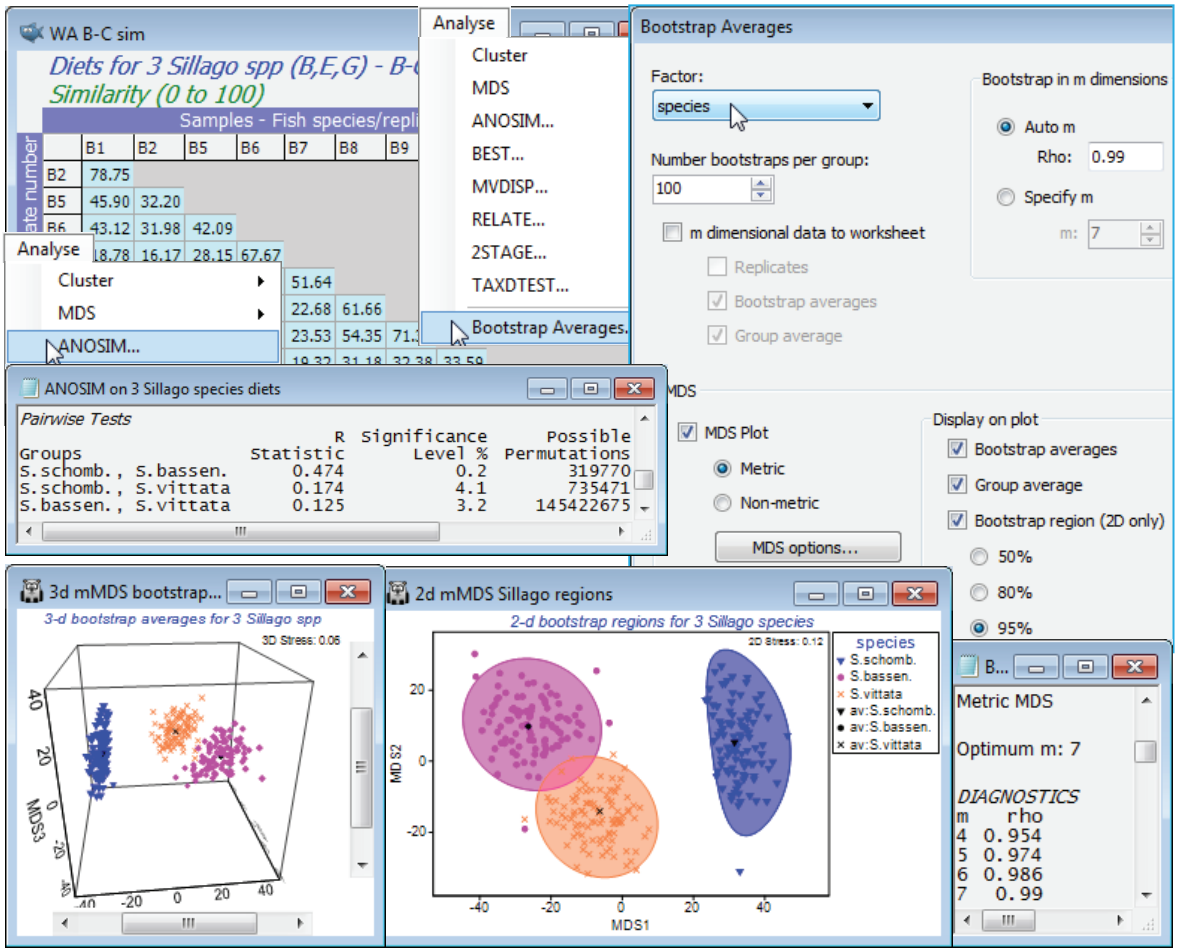
v7



Running the Bootstrap Averages routine

v7

None of the above is necessary in order to create the means plot with regions based on bootstrap averages – it was included purely to note the initial steps the routine takes, under the automatic m option, in order to determine the *m*MDS dimensionality in which bootstrapping will take place. The Pearson ρ values for increasing m = 4, 5, 6, … are sent to the results window, until they reach the given threshold (default ρ≥0.99) or m gets to 10. These ρ values will be displayed before starting the rest of the routine, namely the compute-intensive *m*MDS iterations for large numbers of points. So, run **Analyse>Bootstrap Averages**>(Factor: species) & (Number bootstraps per group: 100) & (•Auto m>Rho: 0.99) & (✓MDS plot•Metric) & (✓Bootstrap averages) & (✓Group average) & (✓Bootstrap region(2D only)•95%). There is also an option to send the m-dimensional data to a worksheet which is not the default and which you do not need to take here. This would allow the bootstrap averages and overall group averages to be placed together with the original replicate-level data (all in the reduced m dimensional space) for a further *m*MDS ordination. However, this will not usually be helpful – the original replicates will, in most cases, already fit poorly into low-d *m*MDS space, so compounding this by adding in several hundred points of bootstrap averages will only make matters worse, and an acceptable stress to allow the viewing of both replicates and a measure of uncertainty in the group means in a single ordination is rather unlikely. This should not be seen in too negative a light since, though the analogue of such a plot is perfectly feasible for univariate data, it is not seen very often there either!



v7 As implied earlier, bootstrap averages are therefore calculated in m = 7 dimensional *m*MDS space, since the ρ value has reached 0.99 at that point (ρ is more or less guaranteed to increase as m gets larger, and could only not do so if a rather sub-optimal ordination has been generated for one of the trial values of m). The MDS section of the Bootstrap Averages dialog then determines what is done with the 300 averages (100 for each group) and the (Euclidean) distance matrix from their 7-d co-ordinate space. Clicking the **MDS options** button leads to the usual *m*MDS dialog and by default (as here) this will produce both a 2-d and 3-d *m*MDS ordination plot from these distances, the 2-d plot displaying the smoothed, nominal 95% regions described earlier (there is no option available for smoothed 3-d regions in the 3-d ordination). Note that the stress is much lower for both plots (0.12 for the 2-d and 0.06 for the 3-d) than for the original replicate-level space, even though there are 300 points rather than the original 38, primarily because these are plots of means of between 8 and 16 replicates and therefore have much lower sampling variability. The structure is also now (inevitably) much simpler, with three fairly well-defined groups of points.

There are several relevant points to note from these plots:

v7 a) the smoothed regions, though they have to be convex, are not just ellipses – the back-transform from ellipses fitted in a transformed space gives good flexibility to mould the regions to the shape of the clusters of bootstrap averages;

b) the bias correction applied at this point tends to make the smoothed envelopes just slightly larger than the observed spread of 95 of the 100 bootstrap averages, though the effect is marginal, with the reasonably large *n* values (8 to 16) for the original replicates;

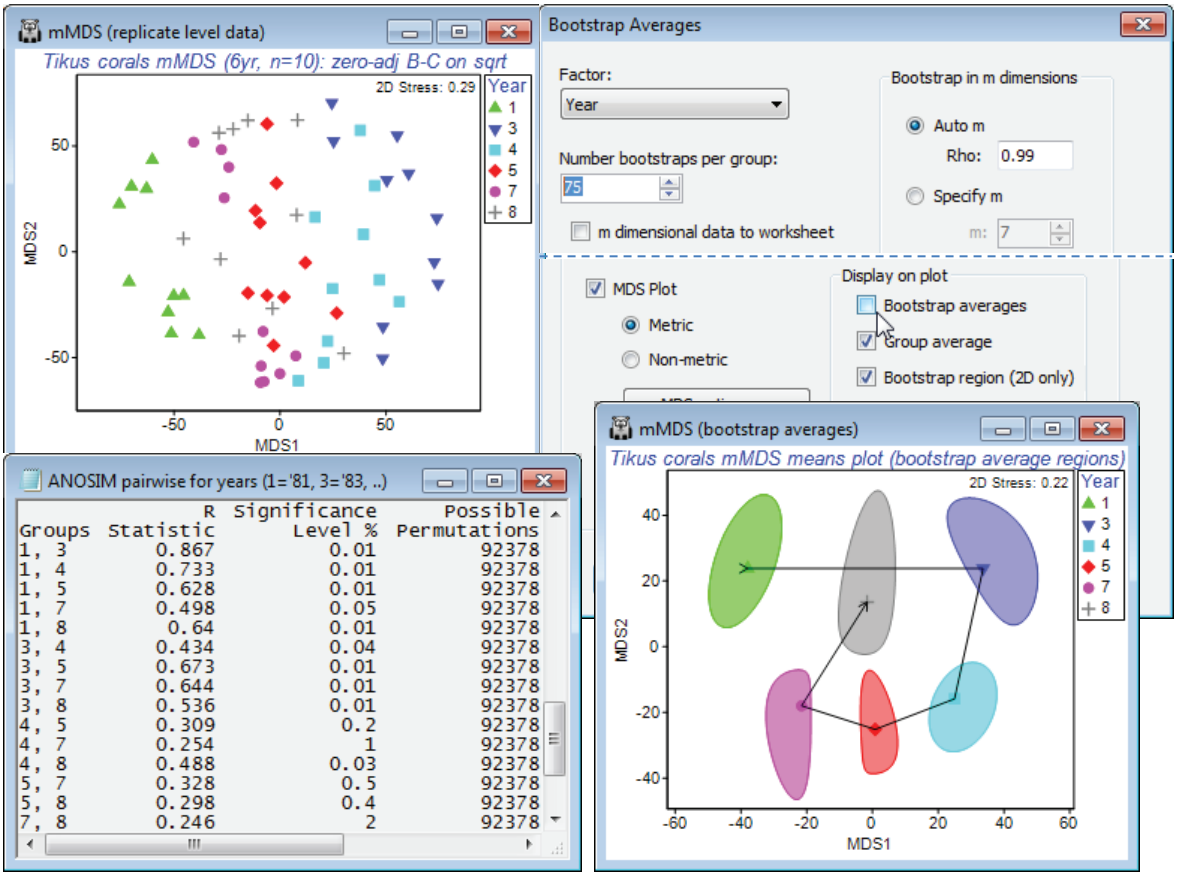
c) the spread of all three clusters of averages is relatively comparable but this is accidental and does not follow from any assumption of constant dispersion. The original replicate-level *m*MDS (though it needs to be interpreted cautiously with stress of 0.22) shows the blue triangles of *S. schombergkii* to have a somewhat less variable diet than the other two species, but it also has half the number of replicates (8, compared with 14 and 16) and therefore the variability will not decrease as much, as a result of the averaging – these two effects tend to cancel each other out;

d) the smoothed regions for *S. bassensis* and *S. vittata* marginally overlap, but in the 3-d solution, with its lower stress, the clusters seem disjunct (though abutting). However, you must resist the

v7 temptation to turn this into a hypothesis test. It is not even true in the simple univariate case (with full normality and constant variance assumptions), that if 95% confidence intervals overlap then their means are not significantly different – and here we have taken some trouble to point out the approximate nature (in several ways) of these nominal 95% envelopes. The formal tests here are those of ANOSIM – and PERMANOVA, since that works with the measurement scale of the dissimilarities (as *m*MDS does) rather than the rank values in ANOSIM. Here, as often, the tests give very similar results, with borderline significance for this pairwise difference (p<3% for both tests).

Bootstrap regions for Tikus coral reef study

The above study was not an example given in Chapter 18 of CiMC and was therefore discussed in detail, but bootstrap average regions are given and interpreted for three other data sets there, and we shall end just by showing a region plot from one of those, for the 6 years (1981, 83, 84, 85, 87 and 88) of coral assemblage data from 10 replicate transects per year at Tikus Island, Indonesia. The workspace Tikus ws was saved in Section 5, where it was used to illustrate the zero-adjusted Bray-Curtis similarity coefficient (see also the analyses for this data in Chapter 16, CiMC). If not available, open file Tikus coral cover from C:\Examples v7\Tikus corals, square root transform it and **Analyse>Resemblance>**(Measure•Bray-Curtis similarity)&(✓Add dummy variable>Value:1), calling it B-C adj. An *m*MDS of the replicate-level data (60 points) has high stress of 0.29 but does seem to show a major change between 81 and 83 (spanning a major coral bleaching event) and a (partial) reversion – the nMDS has a similar pattern but also a high stress (of 0.21). On B-C adj, run **Analyse>Bootstrap Averages>**(Factor: Year) & (Number bootstraps per group: 75) & (•Auto m> Rho: 0.99) & (✓MDS plot•Metric) & (✓Group average) & (✓Bootstrap region•95%) but uncheck the (✓Bootstrap averages) box. The bootstrap averages will still be calculated of course, and used to structure the 2-d *m*MDS space for the display of regions (and the computation time will be non-negligible for an MDS of 450 bootstrap average points!) but there may occasionally be merit in showing just the smoothed regions, with group averages joined, as in Fig. 18.2c of CiMC. It is wise though to run again with the bootstraps displayed, to check the shape of the envelopes against the clusters –1987 shows some non-convexity, also suggested by the position of the group average in relation to the smoothed region. To join the group averages in year order, take **Graph>Special> Overlays>**(✓Overlay trajectory)>(Trajectory numeric factor: Year). Testing by ANOSIM (or by PERMANOVA) does show significant differences between all pairs of years, which permits a clear interpretation of the temporal pattern, in this *means plot* with bootstrap-derived regions.

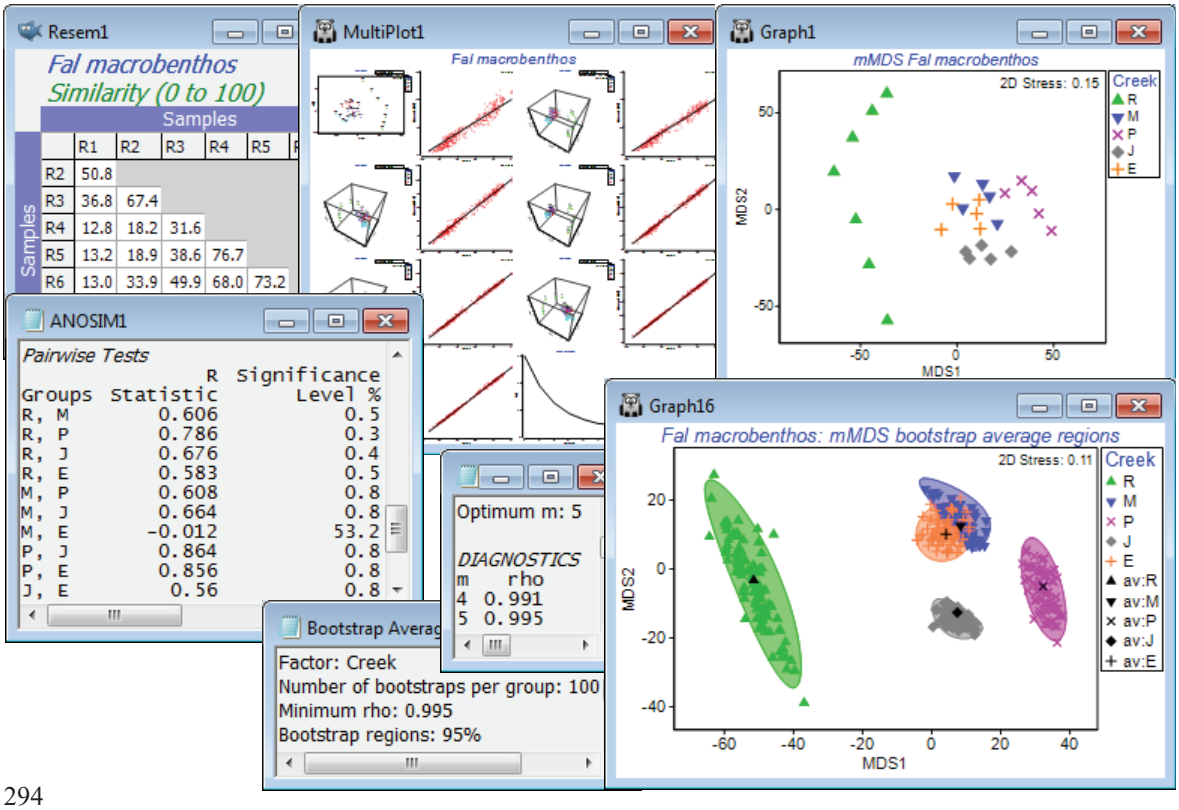


Bootstrap regions for Fal estuary macrofauna

A final example of bootstrap regions which do strongly overlap, and for which the hypothesis tests (such as ANOSIM) give no indication at all that the groups differ, is shown in Fig. 18.7 of CiMC. It can be reproduced here by opening the Fal macrofauna counts data file, in C:\Examples v7\Fal benthic fauna, into either a new workspace or Fal ws saved earlier (in which only the Fal copepod data was examined). The suite of data from these 27 locations from 5 creeks running into the Fal estuary, Cornwall, UK, were introduced in Section 4, and interest is in whether the differing levels of heavy metals in the sediments of these creeks, from historical tin and copper mining in their respective valleys, lead to differing macrofaunal (and meiofaunal) communities in those sediments. Fourth-root transform the Fal macrofauna counts, computing Bray-Curtis similarities and *m*MDS ordination of the replicate-level data. Running 1-way ANOSIM on the 5 creeks, with 7 replicates in Restronguet (R) and 5 for St Just (J), Pill (P), Mylor (M) and Percuil (E), gives strong differences, both in the global test (R = 0.49, p<0.1%) and in all pairwise tests (R>0.55, p<1%) except that between Mylor and Percuil (R = –0.01). Perusal of a means plot is therefore certainly justified.

v7

The stress of 0.15 for the replicate level ordination is quite low for a metric MDS, and if the default range of dimensions requested is changed from 2-3 up to (for example) 2-8, it is clear from Shepard diagrams that by the time m=4 is reached, the m-dimensional *m*MDS distances are a very good fit to the full-dimensional resemblance matrix – in fact the later output of the bootstrap routine shows that this already gives a Pearson correlation of 0.991. Now enter the similarity matrix to **Analyse> Bootstrap Averages**>(Factor: Creek) and increase the number of bootstrap averages from the suggested default of 60 to nearer 100, depending on the speed of your machine. The (•Auto m) choice, with a ρ>0.99 threshold, as indicated, does lead to bootstrapping in m = 4 dimensions. You might wish instead to experiment with (•Specify m) at a higher, fixed level of 5 or 6, or increase the threshold in the automatic routine to ρ>0.995 or 0.999, but it will make negligible difference to the outcome in relation to the variation from run to run of the same m, resulting from the random differences in the bootstrap samples selected. (It is always a good exercise to repeat the bootstrap routine under the same conditions, and will discourage you from over-interpreting the minutiae of the region shapes!). Though the below did use 100 bootstraps from each creek, a replication level of only *n*=5 in four of the creeks must be considered absolutely minimal, and the striations in the bootstrap average points, which are just discernible in the plot, result from the fact that there are then only a possible 126 bootstrap averages (not equally likely) and several will have been created more than once. (The *n*=7 for Restronguet gives 1716 possibilities and thus more of a continuum of average values). So the plots should again be interpreted with caution, but the pattern of differences among creeks is clear, and fully consistent with the hypothesis testing.



Index to data sets

Pages on which the specified data sets are analysed; **bold** indicates location of an introduction to the dataset and a source reference. The list is ordered by first appearance and the marginal box gives the data directory.

Acknowledgements

We thank our many collaborators, correspondents and workshop participants from around the world, far too numerous to list, for their continued enthusiasm for the PRIMER software and suggestions for improvement, some of which we hope have been incorporated in this latest incarnation of the routines. In terms of the new methods in PRIMER 7, special thanks are due to Paul Somerfield (Plymouth Marine Laboratory, UK) and Marti Anderson (Massey University, Auckland, NZ), along with several colleagues/collaborators at Murdoch University, Western Australia (Fiona Valesini, James Tweedley, Margaret Platell, Ian Potter and Richard Warwick). KRC also acknowledges his honorary fellowship at Plymouth Marine Laboratory and his adjunct professorship at Murdoch University. We are also keen to thank the many individuals who have, through undertaking the local organisation of PRIMER training workshops, allowed us to get valuable feedback on the ways in which PRIMER is being used, and how it might be more effective. Special mention needs to be made of those who have repeatedly volunteered (!) over the last 15 years: Gerhard Pohle (Huntsman MSC, St Andrews NB, Canada; 8 times), Euan Harvey (formerly UWA now Curtin University, W Australia; 7), Steve Murray (CSU Fullerton CA, USA; 6), Walt Jaap (USF St Petersburg FL, USA; 6), Rachel Gorman (La Trobe University, Wodonga VIC, Australia; 5), and staff of the Marine Biological Association, Plymouth, UK (~25); and those who have led PRIMER (and PERMANOVA+) workshops with us, and for us, over this time, primarily Paul Somerfield and Marti Anderson, but also JJ Cruz and Edlin Guerra (U Simón Bolívar, Caracas & U Oriente, Isla Margarita, Venezuela) and Victor Quintino (U Aveiro, Portugal).

Bob Clarke & Ray Gorley (2015)