10. Centroid plots

10.1 Ordinations for multi-factor designs

Rationale

When considering the response of a whole set of variables (such as the abundances of species or taxa) simultaneously to a suite of several factors (e.g., arising from a multi-factor experiment or sampling design), it can be difficult to visualise salient structures and patterns in the data. One common problem is that multi-factor designs, when appropriately replicated, can yield a large total number of sampling units. A non-metric (or metric) multi-dimensional scaling (MDS) ordination done on a large number of samples can be very difficult to interpret. First, the 2D (and even 3D) stress might be quite high (> 0.20), precluding interpretability. Second, the residual variation (i.e., variation among the sampling units within each cell of the study design) is often quite large, and can mask essential patterns happening across the main factors of interest.

In univariate analyses of groups of samples (e.g., as in an ANOVA), one would commonly examine plots of means, rather than plots of raw sample values, to visualise patterns. In a similar way, for multivariate analyses, it is very useful to be able to visualise *distances among the centroids* in the space of a chosen resemblance measure. We can usefully construct ordinations from distance matrices among centroids that have been constructed from:

- levels of factors that are the main effects (*main effects plots*);
- combinations of levels of factors that are crossed with one another (*interaction plots*).

Ordination plots of distances among centroids were described by Anderson (2017) . In PRIMER 7 , it was possible to calculate distances among centroids, based on any given grouping factor (using **PERMANOVA+** > **Distance Among Centroids...**). One can use this tool to generate interaction plots of interest by first creating factors that consist of combinations of levels of some chosen factors (e.g., using **Edit** > **Factors...** > **Combine...**).

In PRIMER 8, one can generate resemblance matrices and ordination plots of either (i) main-effect centroids or (ii) interaction centroids, automatically, by reference to a specific Design file. We shall outline briefly here (below) how resemblance matrices among centroids are constructed. We will then demonstrate these new practical tools and their utility for visualising and interpreting salient patterns in a multi-factor design by way of an example.

Distances among centroids

Let ${\bf Y}$ be a matrix of N rows (sampling units) by p columns (variables). Let ${\bf D} = \lbrace d_{ij} \rbrace$, $i = 1, ..., N$; $j = 1, ..., N$ be the distances or dissimilarities between every pair (i,j) of sampling units. If ${\bf D}$ contains Euclidean distances, then the distances among centroids are equivalent to Euclidean distances among the arithmetic averages calculated

separately for each variable. This equivalence does not hold, however, for non-Euclidean dissimilarities. Distances among centroids based on some other chosen dissimilarity measure (such as Bray-Curtis) are calculated as follows:

Step 1 - Calculate Gower's ${\bf G}$ matrix from ${\bf D}$

As in Gower (1966) , obtain $(N \times N)$ matrix $\bf G$ by first defining matrix ${\bf A} = \lbrace a_{ij} \rbrace = \lbrace - 0.5 \cdot d_{ij}^2 \rbrace$, then centring the elements of this matrix by its row-means, $\bar{a}_ {i\cdot}$, its column-means, $\bar{a}_ {\cdot j}$, and its overall mean, $\bar{a}_ {\cdot \cdot}$, to yield ${\bf G}$, i.e.,

${\bf G} = \lbrace g_{ij} \rbrace = \lbrace a_{ij} - \bar{a}_ {i\cdot} - \bar{a}_ {\cdot j} + \bar{a}_ {\cdot \cdot} \rbrace$

Step 2 - Obtain the full set of principal coordinate (PCO) axes from matrix ${\bf G}$
This is done by performing an eigenvalue decomposition of matrix ${\bf G}$. The resulting eigenvectors are each standardised by the absolute value of their respective eigenvalue. At this step, it is important to keep track of those eigenvectors that are associated with *positive* eigenvalues, and those that are associated with *negative* eigenvalues (if any), as two separate sets.

Step 3 - Calculate centroids as averages along PCO axes
Suppose there are $\ell = 1, ..., c$ cells (or specified groups of sampling units), and we require a centroid for each of these. The centroids are obtained as the arithmetic averages of the sampling units belonging to each cell (or group), calculated separately along each PCO axis.

Step 4 - Calculate distances among centroids
For every pair of centroids (ℓ, ℓ'), $\ell = 1, ..., c$ and $\ell' = 1, ..., c$, calculate Euclidean distances separately in each of two sets: one based on PCO axes corresponding to non-negative eigenvalues $(d_{\ell \ell'}^+)$ and one based on those corresponding to negative eigenvalues $(d_{\ell \ell'}^-)$, if any. Next, the $(c \times c)$ matrix of distances among centroids in the space of the chosen dissimilarity measure is then:

${\bf D}^{\left[C \right]} = \lbrace d_{\ell \ell'}^{\left[C \right]} \rbrace$, where $d_{\ell \ell'}^{\left[C \right]} = \sqrt{ | (d_{\ell \ell'}^+)^2 - (d_{\ell \ell'}^-)^2 } |$

It is worth noting here that the distances among centroids can also be calculated directly from matrix ${\bf G}$, without using PCO axes. See section 5.1 in Anderson (2017) for details.

10.2 Main effects plot

What is a 'main effects plot'?

In a main effects plot, we calculate and then show in an ordination diagram *a centroid for each of the levels of each factor listed in the design file*.¶ We may also (optionally) show the overall centroid as well. The centroids, and distances/dissimilarities among them, are calculated in the full high-dimensional space of a chosen resemblance measure.

A two-way crossed example

Let's consider a study by Glasby (1999) on the development of subtidal epibiotic assemblages. It was proposed that differences in two factors: (i) shading and (ii) proximity to the seafloor, could explain previously observed differences between assemblages of sessile organisms on rocky reefs *vs* pier pilings in sheltered embayments of Sydney Harbour, Australia. Four replicate sandstone settlement plates (15 cm $\times$ 15 cm) were placed in an embayment in Middle Harbour (a branch of Sydney Harbour) in each of 3 shading treatments: 'Shade' (shaded surfaces with an opaque Perspex roof), 'Control' (a procedural control with a clear Perspex roof), and 'Open' (surfaces without a roof) in each of 2 positions relative to the seafloor ('Near' to and 'Far' from the seafloor); see Fig. 10.1. The percentage cover values of p = 46 taxa colonising the settlement plates were recorded after 33 weeks of deployment.

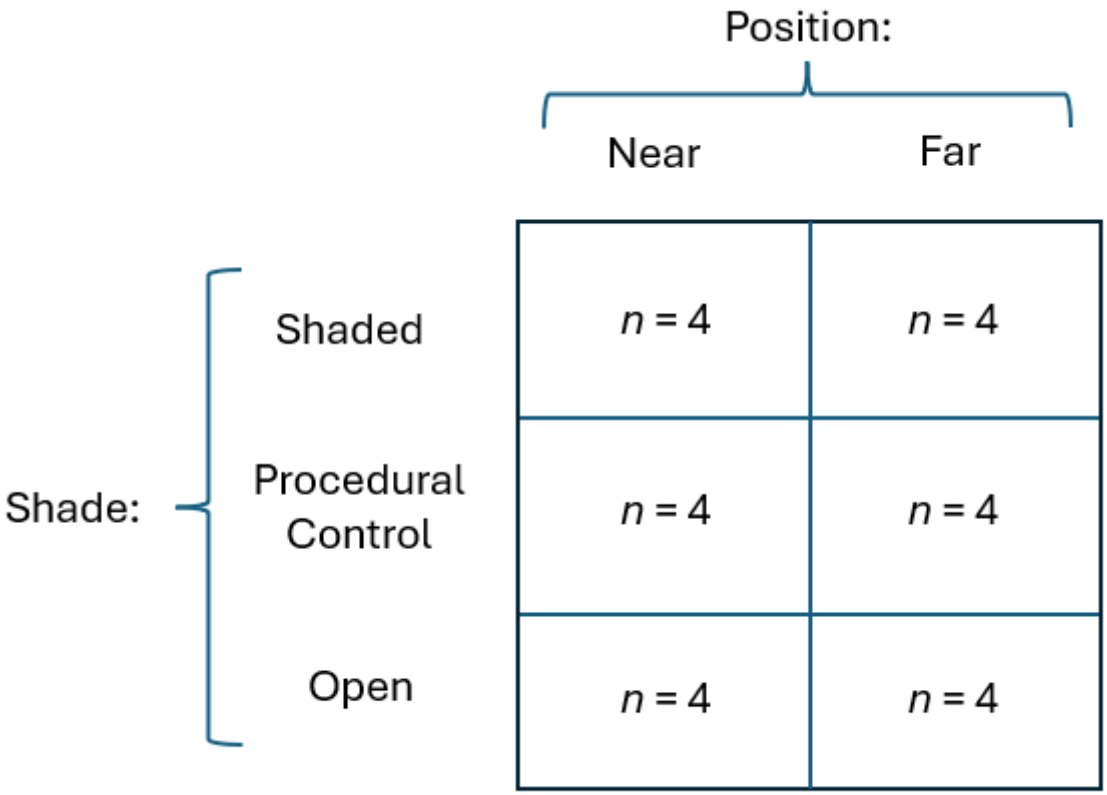


Fig. 10.1. Schematic diagram of the two-factor crossed study design described by Glasby (1999) .

The data for this example can be found in the file 'Sydney_subtidal_epibiota.pri' in the 'Examples_P8' > 'Subtidal_epibiota' folder. A non-metric MDS ordination of the individual sampling units on the basis of the Bray-Curtis resemblance measure, calculated after applying a square-root transformation, is shown in Fig. 10.2 below.

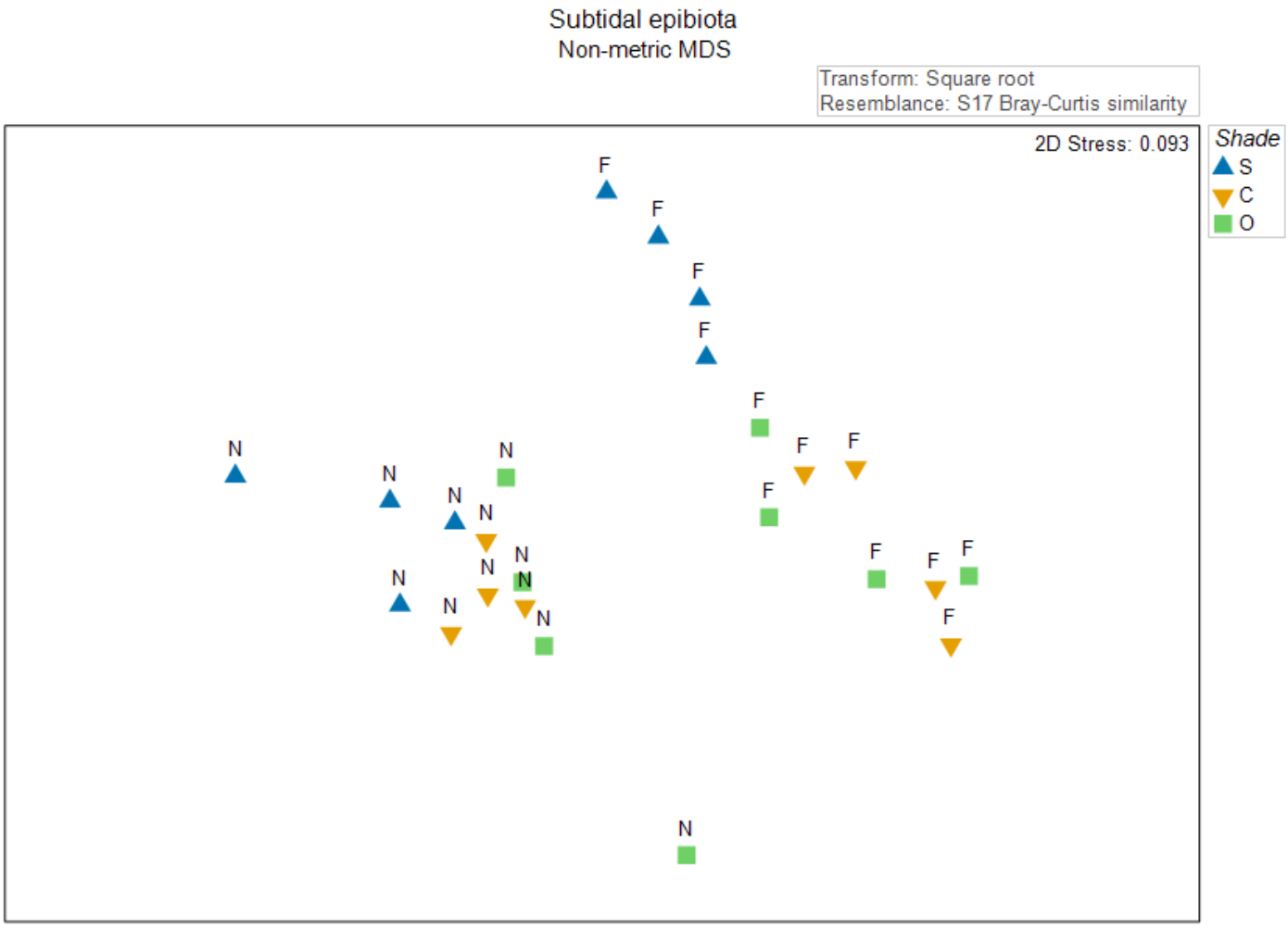


Fig. 10.2. Non-metric MDS of the individual sampling units from the two-factor crossed study design described by Glasby (1999) . Symbols correspond to the factor of 'Shade' ('S' = Shade, 'C' = Procedural Control, 'O' = Open surfaces), and labels correspond to the factor of 'Position' ('N' = Near, 'F' = Far).

It is useful to think about where the main effect centroids would be in this diagram, even though we know that there is some stress, and so their true positions by reference to the full high-dimensional system are not able to be represented perfectly here. Let's suppose, just for the moment, that all of the variation is captured in these two nMDS axes. In that case, we can calculate the arithmetic averages along each of the nMDS axes to get the positions of the overall centroid and the centroids for the levels of each of the main effects. If we plot these 'main effect' centroids in the diagram, along with the replicates, we have Fig. 10.3.

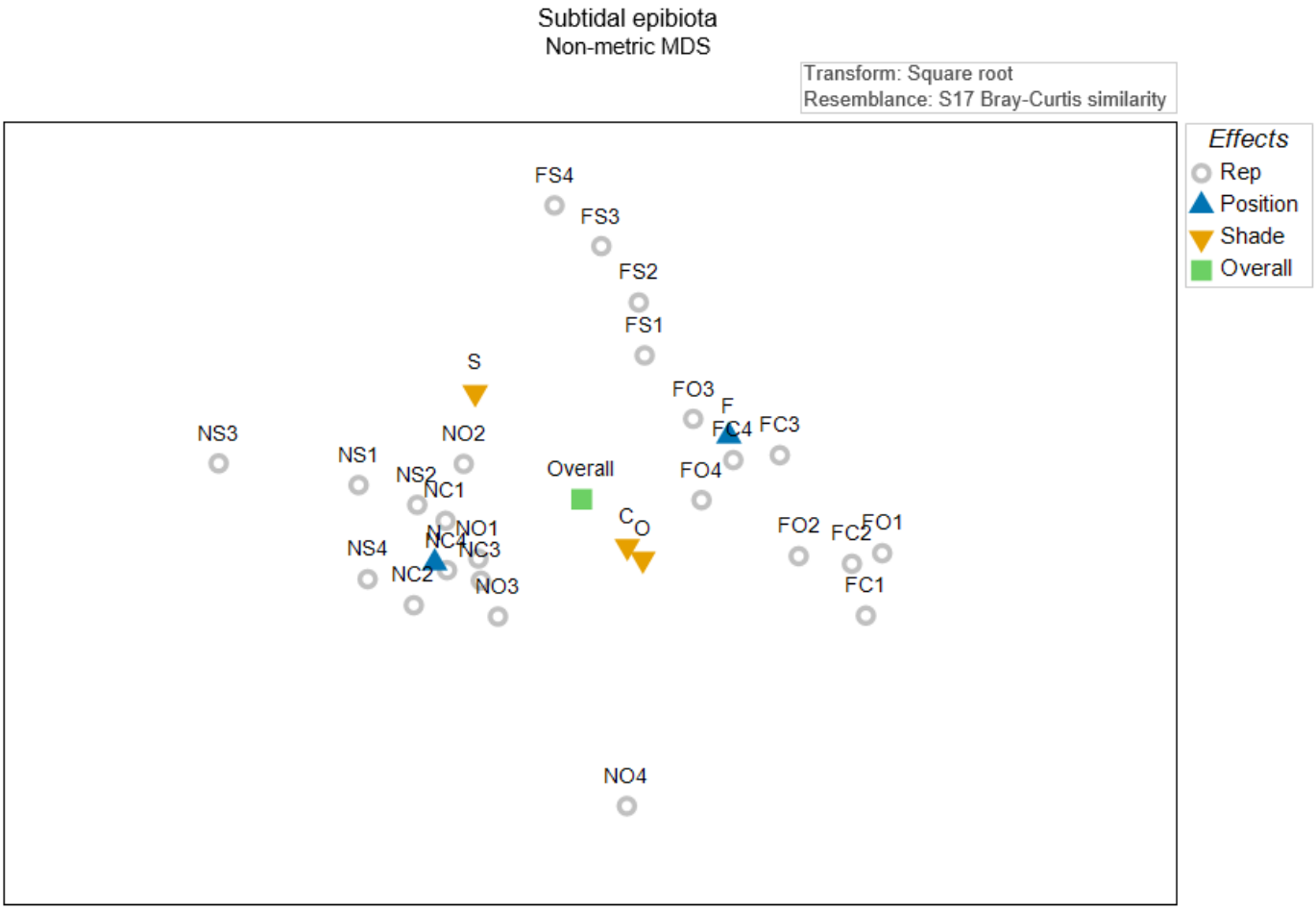


Fig. 10.3. Non-metric MDS of the individual sampling units (grey circles) from the two-factor crossed study design described by Glasby (1999) , along with the overall centroid (in green), the centroids for the position treatments (N and F, in blue), and the centroids for the shade treatments (S, C and O, shown in amber). All of these centroids were calculated 'post hoc', simply as the arithmetic averages in this 2D nMDS space.

This plot is somewhat useful, but what we really want, in fact, is to calculate these centroids in the original space of the resemblance measure (and not on the basis of the 2D Euclidean space of nMDS axes). This can be done using the method described in section 10.1. We can construct the distances among these main effect centroids, then perform an ordination to create a 'main effects plot' (Fig. 10.4).

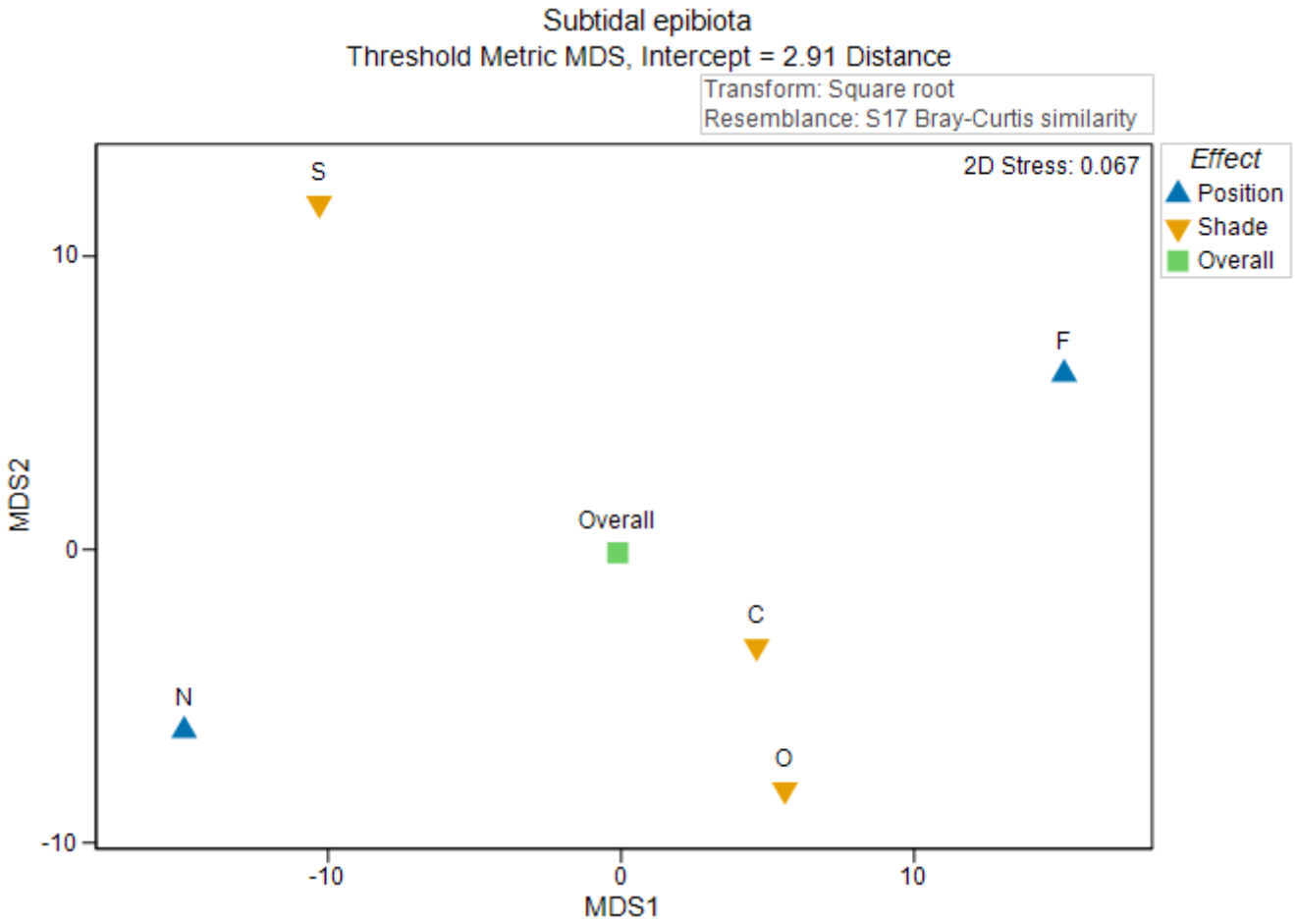


Fig. 10.4. Threshold metric MDS plot of the centroids for the main effects in the two-factor crossed study design described by Glasby (1999) .

As there are (typically) far fewer points in an ordination plot of distances among centroids, we tend to be able to get an interpretable diagram with tolerably low stress (< 0.2) using metric or threshold metric MDS; we typically do not need to use a (more forgiving) non-metric MDS to visualise these relationships. This is clearly advantageous, as our resulting ordination plot will therefore have labels on its axes, hence the relative sizes of effects (in the units of the original chosen resemblance measure) can be visualised, quantified and compared with one another. In the present example, the effects of 'Position' (correlated somewhat with the first tmMDS axis) are somewhat larger in size than the effects of 'Shade' (correlated somewhat with the second tmMDS axis).

Multivariate dissimilarity-based effects

Recall that an 'effect', in a univariate ANOVA context, can be described as the deviation of a group mean from the overall mean. As the axes for main effect plots are in units of the chosen dissimilarity measure (Bray-Curtis, in this case) we are able to get a real sense of the multivariate effect sizes as 'deviations of group centroids from the overall centroid', just as PERMANOVA would measure such effects in its partitioning of the full resemblance space (Fig. 10.5).

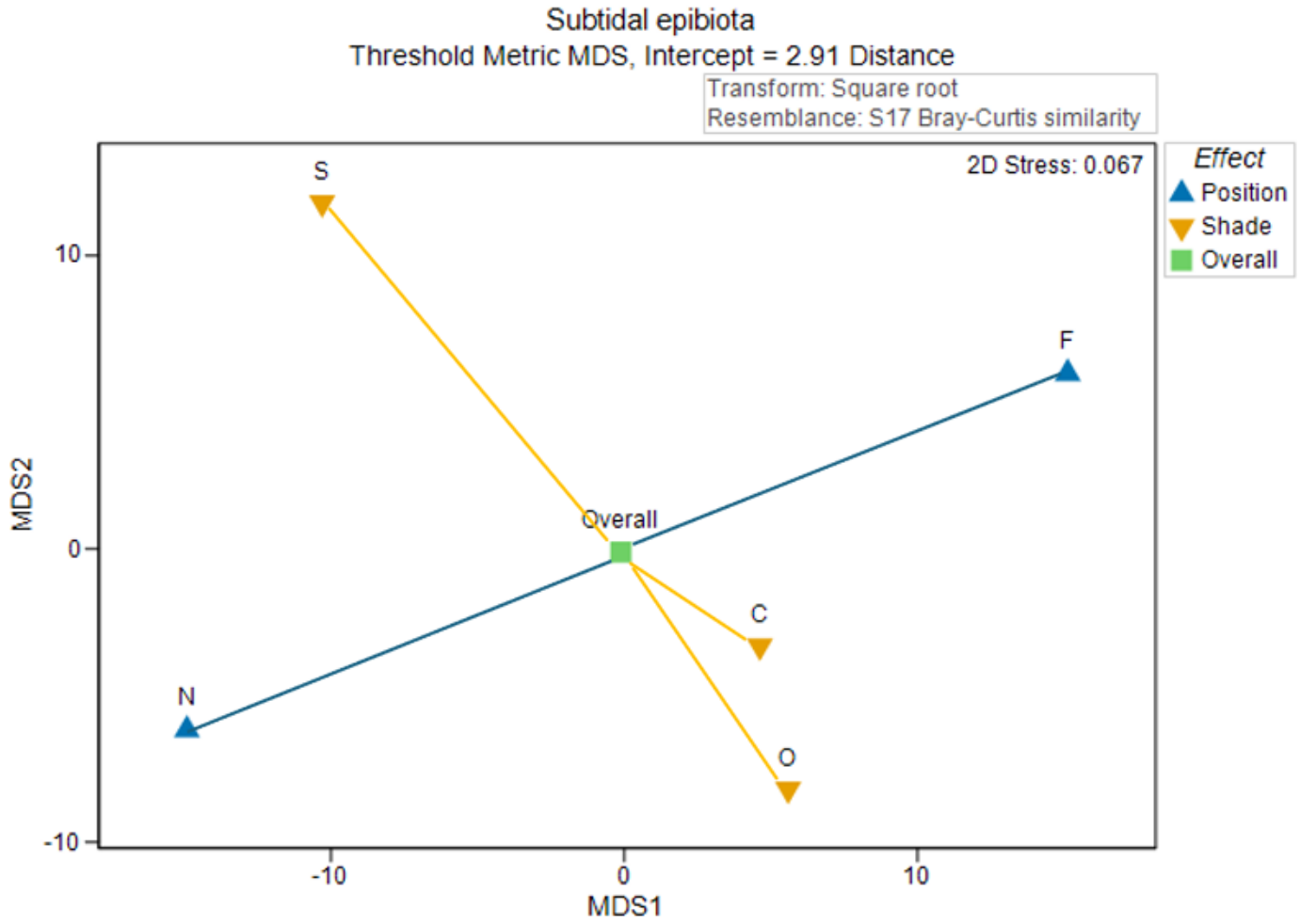


Fig. 10.5. Threshold metric MDS plot of the centroids for the main effects in the two-factor crossed study design, as shown in Fig. 10.4, including lines that estimate the sizes of effects as PERMANOVA would measure them. These correspond to 'deviations of centroids from the overall mean' for each factor. Position effects are shown in blue; Shade effects are shown in amber.

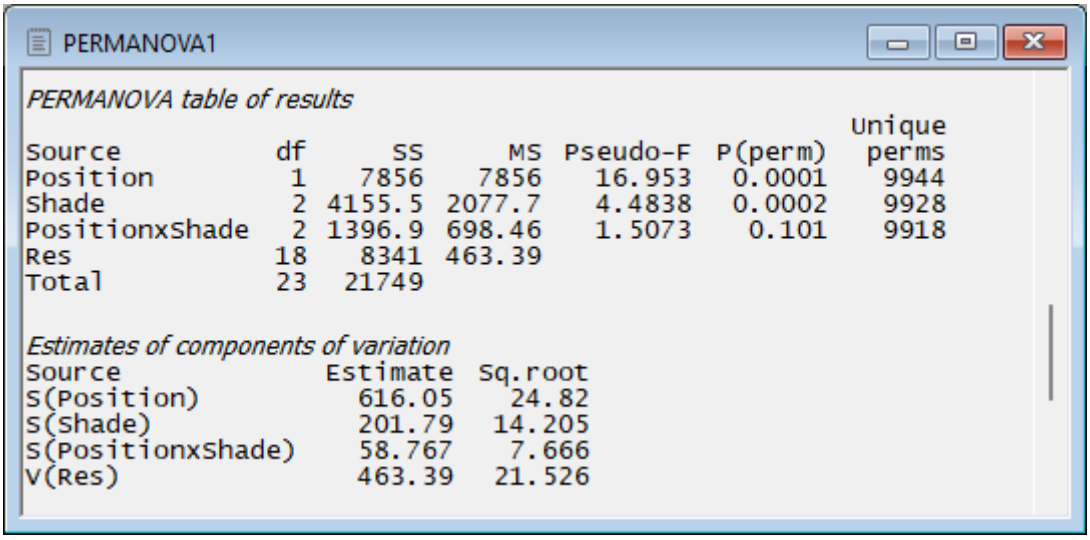
Note that the intercept from the threshold metric MDS is shown in the subtitle of Fig. 10.5. The intercept from the tmMDS Shepard diagram is interpretable as the minimum dissimilarity in the original multivariate space between any two points that occupy the same position in the tmMDS plot (i.e., that have a distance between them on the plot of zero). Therefore, this intercept value should be *added on* to any inter-point distances measured or estimated from the tmMDS ordination plot.[‡]

In the present example, the effects of 'Position' (correlated somewhat with the first tmMDS axis) are somewhat larger in size than the effects of 'Shade' (correlated somewhat with the second tmMDS axis). Also, we see that the effect of being an assemblage colonising a panel 'Near' the seafloor generates a deviation from the overall centroid of about 15-20 units (in Bray-Curtis space), in a direction towards the bottom left of the diagram. The effect of being 'Far' from the seafloor is a deviation that is equal in size to this, but in the opposite direction.[†] We can also see that the effect of being in a 'Shade' treatment shifts assemblages a rather similar magnitude away from the overall centroid (i.e., a distance of approximately 15-20 units in Bray-Curtis space), but in a completely different direction (i.e., towards the top of the diagram in this particular case). Also, it is clear that assemblages colonising surfaces in 'Open' and 'Control' treatments are not too dissimilar

from one another, with their centroids differing by something that is likely to be less than 10 units in the full Bray-Curtis space.

Consider alongside output from a PERMANOVA partitioning

If we next do a PERMANOVA paritioning on the basis of the Bray-Curtis resemblance measure, after a square-root transformation, we see the following results:


PERMANOVA1

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
Position	1	7856	7856	16.953	0.0001	9944
Shade	2	4155.5	2077.7	4.4838	0.0002	9928
PositionxShade	2	1396.9	698.46	1.5073	0.101	9918
Res	18	8341	463.39			
Total	23	21749				

Estimates of components of variation

Source	Estimate	Sq.root
S(Position)	616.05	24.82
S(Shade)	201.79	14.205
S(PositionxShade)	58.767	7.666
V(Res)	463.39	21.526

This output provides not only the PERMANOVA table of results, but also direct estimates of the sizes of effects for each term in the model in the section entitled '*Estimates of components of variation*'. The column labeled '*Estimate*' is interpretable as the sum of squared fixed effects (divided by degrees of freedom) in the full Bray-Curtis space (in the case of fixed factors, as in this example). The square root of this value (in the column labeled '*Sq.root*') is therefore interpretable as a type of 'standard deviation' attributable to that factor in that space. Thus, these '*Sq.root*' values should correspond well with the mean deviations from the overall centroid for a given factor.

The key point here is that the main effects plot can be examined alongside the '*Estimates of components of variation*' section of the PERMANOVA output file in order to guage the size and the relative importance of the factors in explaining overall variation. The plot, furthermore, shows the positions of the groups in the high-dimensional space relative to one another, which can also be quite useful. Such positions may not be discernable in ordinations of individual replicates from multi-factor designs where residual variation is high.

A few cautionary notes regarding interpretation

1. **Each centroid should really be accompanied by some sort of measure of its variability.** Measures of variation for these centroids are *not* included in these main effects plots. The main effects plots offered here are designed to permit clarity in interpreting effect sizes and the relative positions of centroids due to different factors in high-dimensional multi-factor designs.

   If we had a distribution of sampling units (in a p-dimensional Bray-Curtis space, say),

that could be treated as multivariate normal (MVN), having a mean parameter vector ${\bf \mu}$ and a variance-covariance parameter matrix ${\bf \Sigma}$, then under the multivariate central limit theorem, the distribution of a centroid calculated from a sample of size N from that distribution will converge to a MVN distribution with a mean parameter vector ${\bf \mu}$ and a variance-covariance parameter matrix of ${\bf \Sigma}/N$. The multivariate central limit theorem also holds for non-multivariate-normal distributions. Thus, we should expect (all else being equal) that the variability of a centroid that was calculated from a larger number of replicates will be less than the variability of a centroid that was calculated from a smaller number of replicates.

It would be nice to show these differential measures of variability on a main effects plot somehow. However, we are still stuck with the fact that the dimensionality of the system is likely to be too high (with p often being close to or even greater then N) for us to feel comfortable estimating all of the parameters in matrix ${\bf \Sigma}/N$. Some type of bootstrap or jacknife approach might be used to advantage here, but this has not yet been implemented for these plot types in PRIMER (yet).

2. **The greater the number of samples used to calculate a given centroid, the less variable we expect those centroids to be.** This practical point should not be glossed over. There are actually two things going on here. One is driven by the multivariate central limit theorem. Clearly the variation in centroids (${\bf \Sigma}/N$) will decrease with increases in N.

   In addition to this, however, we need to remember that the species-area relationship operates almost ubiquitously in the majority of ecological systems. In other words, the more samples we take, the more species (or taxa) we shall see, overall. This means that centroids calculated from a large number of sampling units will tend to have a greater total richness than centroids calculated from a small number of sampling units. Also, recall that sparser sampling units (or centroids) will tend to look more 'spread out' in an ordination diagram based on the Bray-Curtis (or Jaccard or Sorensen) measure.

   The take-home message from this is that we might expect, *a priori* (all else being equal), that centroids constructed from factors that have many groups (hence where the centroid for each group is calculated using fewer samples) will look more spread out from one another (i.e., may appear to have larger effects in the Bray-Curtis space) relative to factors that have few groups (where the centroid for each group is calculated using a larger number of samples).

   This latter phenomenon is the reason that, should the user choose to include replicates along with the main effect centroids, these will tend to appear all spread out around the edges of the resulting ordination plot. Each replicate will (typically) have fewer species (or taxa) than the centroids will do, so they generally therefore have fewer species in common with one another, yielding lower similarities and greater spread among them.

[¶] *As described in section 10.1, we can calculate these centroids in the space of a chosen resemblance measure by calculating averages along each of the full set of principal coordinate (PCO) axes derived from the dissimilarities, keeping careful track of those axes that correspond to positive vs negative eigenvalues.*

[†] *For any factor that has two groups with equal sample sizes, their effects will be equivalent and in opposite directions in the full multivariate space.*

[‡] *We could alternatively use metric MDS here instead, in which case the intercept is forced to be zero, so no such 'added distance' (threshold) would be necessary. The trade-off here, however, is that metric MDS will always have higher stress than threshold metric MDS. This is why threshold metric MDS is the default method used to create centroid plots in PRIMER, but the user can always choose which MDS flavour they wish in the dialog. If there are a great many centroids to plot, then it is possible that non-metric MDS would be needed (to keep the stress low), but this will, in turn, sacrifice the interpretability of the resulting plot, which will have no axis labels, hence will have no direct quantitative interpretability regarding the sizes of effects - only some indication of their relative sizes.*

10.3 Interaction plot

What is an 'interaction plot'?

Although main effects plots can help us to visualise the main effects of factors and permit us to guage their relative importance in explaining overall variation, centroids based on individual main effects *ignore* all other factors in the study design. However, many systems are *interactive*, and if two (or more) factors do interact with one another, then what we generally want is to visualise how differences in the positions of centroids for a given factor *vary* across levels of one or more other factors. Thus, it is the centroids based on the *cells* in the (crossed) study design that are of interest here.

In an interaction plot, we calculate and then show in an ordination diagram *a centroid for each combination of levels of factors listed in the design file*. The centroids of the combinations of factor levels, and distances/dissimilarities among them, are calculated in the full high-dimensional space of a chosen resemblance measure.

In PRIMER 8, an interaction plot can be obtained directly from the resemblance matrix among replicates for any design up to three factors. Alternatively, an interaction plot can manually be produced for any number of factors *via* the following three steps:

1. From a resemblance matrix, choose **Edit** > **Factors...** > **Combine...** and obtain a factor that consists of the combined levels of two or more factors of your choice.
2. From the resemblance matrix, choose **PERMANOVA+** > **Distances Among Centroids...** and calculate these on the basis of the combined factor you created in step 1.
3. From the resemblance matrix produced at step 2, choose **Analyse** > **MDS** and create an ordination of these combined-factor centroids (using whatever flavour of MDS you wish: mMDS, tmMDS or nMDS).

A two-way example

Veale *et al.* (2014) described a study of nearshore fish assemblages in the Leschenault estuary in Western Australia. For the set of data from this study that we will examine here (found in the file named 'Leschenault_fish_counts.pri, located in the 'Examples_P8 > Leschenault_fish' folder), fish assemblages were sampled using 21.5m seine nets on n = 6 to 8 occasions in each of 4 seasons ('Sp' - Spring, 'S' - Summer, 'A' - Autumn and 'W' - Winter) at each of 4 regions ('B' - Basal, 'L' - Lower, 'U' - Upper and 'A' - Apex) of the estuary (Fig. 10.6).

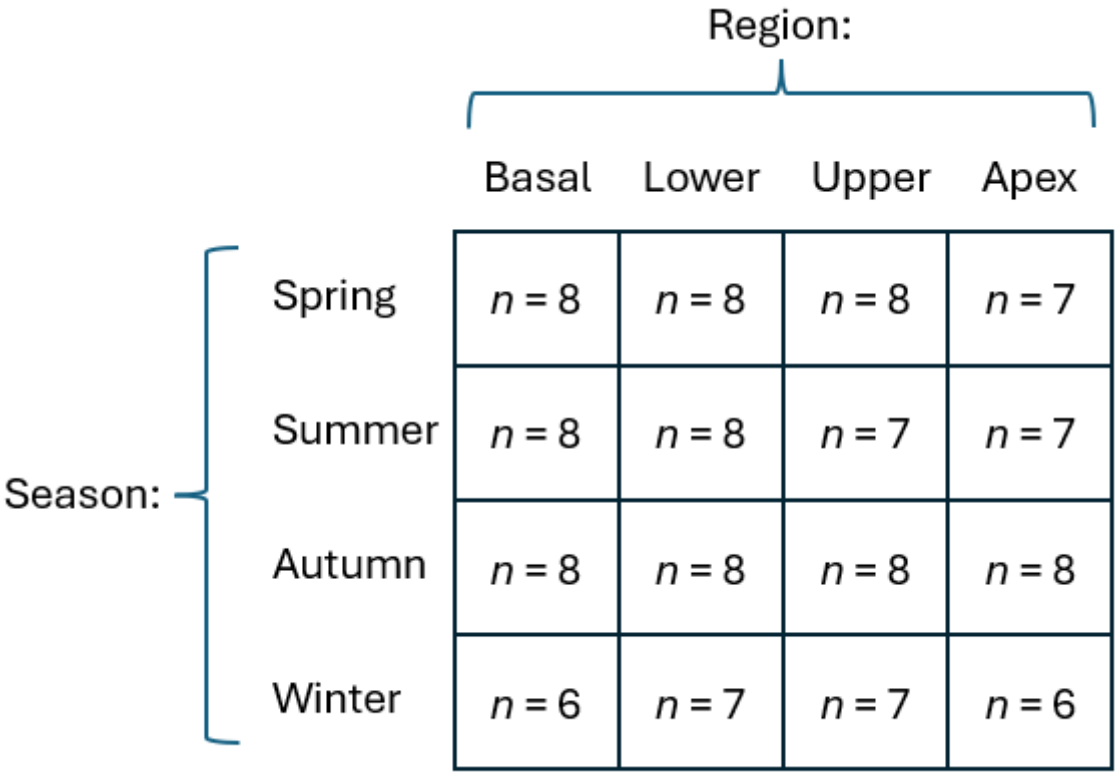


Fig. 10.6. Schematic diagram of a two-factor design, a subset of data from the study decribed by Veale et al. (2014) .

Given that many fish species tend to school (aggregate), a useful pre-treatment option here is to apply *dispersion weighting* (see Clarke *et al.* (2006a)). After applying dispersion weighting (using groups corresponding to the combined factor of Season-by-Region), followed by a square-root transformation, we can calculate Bray-Curtis resemblances and create a non-metric MDS plot among the replicate samples, as shown in Fig. 10.7.

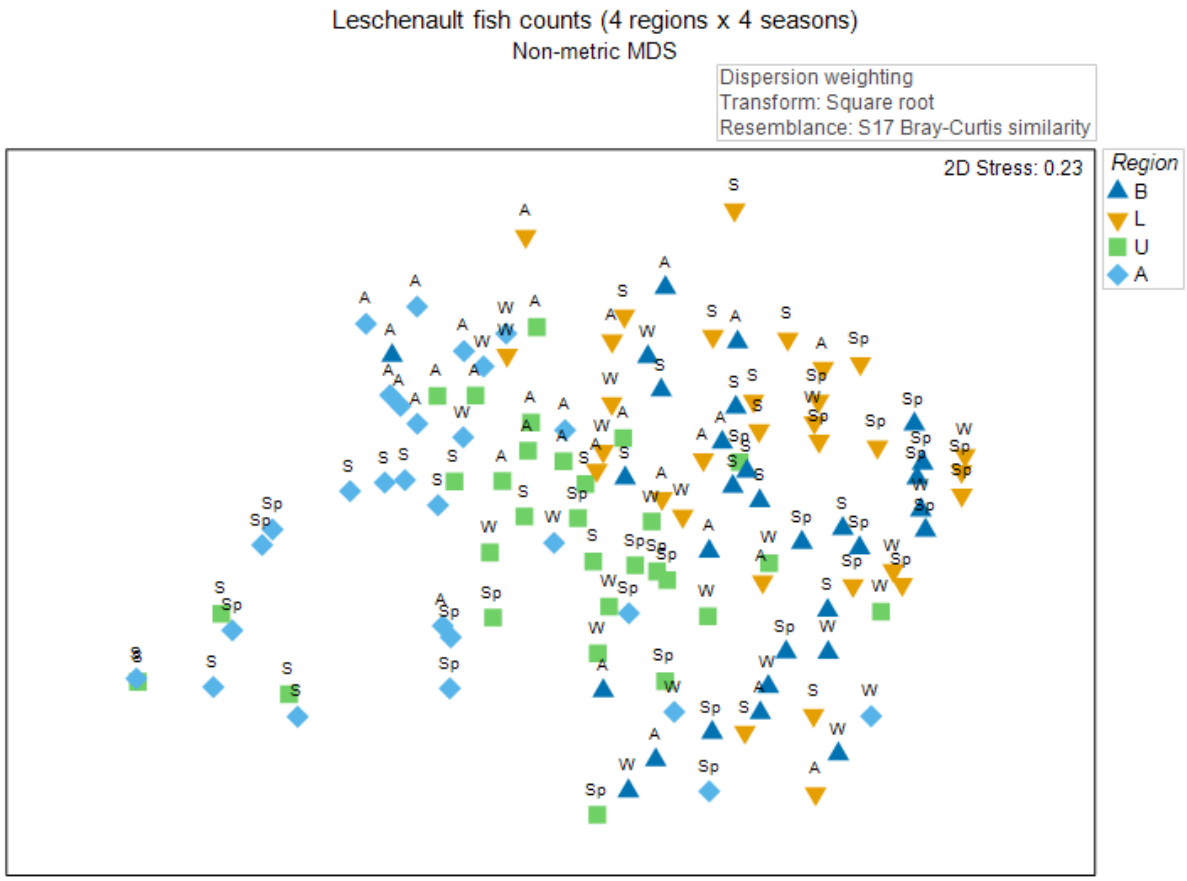


Fig. 10.7. Non-metric MDS of the individual sampling units from a two-way crossed study design of fish assemblages (Veale et al. (2014)). Symbols correspond to the factor of 'Region' ('B' = Basal, 'L' = Lower, 'U' = Upper, 'A' = Apex), and labels correspond to the factor of 'Season' ('Sp' = Spring, 'S' = Summer, 'A' = Autumn, 'W' = Winter).

This nMDS plot is pretty messy. It is difficult to see any clear seasonal or regional patterns in this plot, and the stress is also too high to permit useful interpretation. High stress and a rather messy "dog's breakfast" sort of display is unfortunately a rather typical thing to encounter when we try to create an ordination of a large number of individual samples (here there are 119).

PERMANOVA detects a significant interaction

When we run a PERMANOVA partitioning on this two-factor design (both factors are treated as fixed here) for these data (once again, on dispersion-weighted data that have been square-root transformed and on the basis of the Bray-Curtis resemblance measure), we see the following:

PERMANOVA1

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
Season	3	42986	14329	6.0307	0.0001	9863
Region	3	54372	18124	7.6313	0.0001	9887
SeasonxRegion	9	38834	4314.9	1.8277	0.0001	9769
Res	103	2.4317E+05	2360.9			
Total	118	3.7936E+05				

Estimates of components of variation

Source	Estimate	Sq.root
S(Season)	402.56	20.064
S(Region)	529.95	23.021
S(SeasonxRegion)	263.65	16.237
V(Res)	2360.9	48.589

There is clearly a highly significant interaction between Season and Region in their effects on these fish assemblages ($F_{9,103}$ = 1.83, P = 0.0001). An interaction plot (i.e., an ordination plot of the centroids for all 16 Season$\times$Region combinations of factor levels, calculated in the full high-dimensional space of the chosen resemblance measure) is shown in Fig. 10.8.

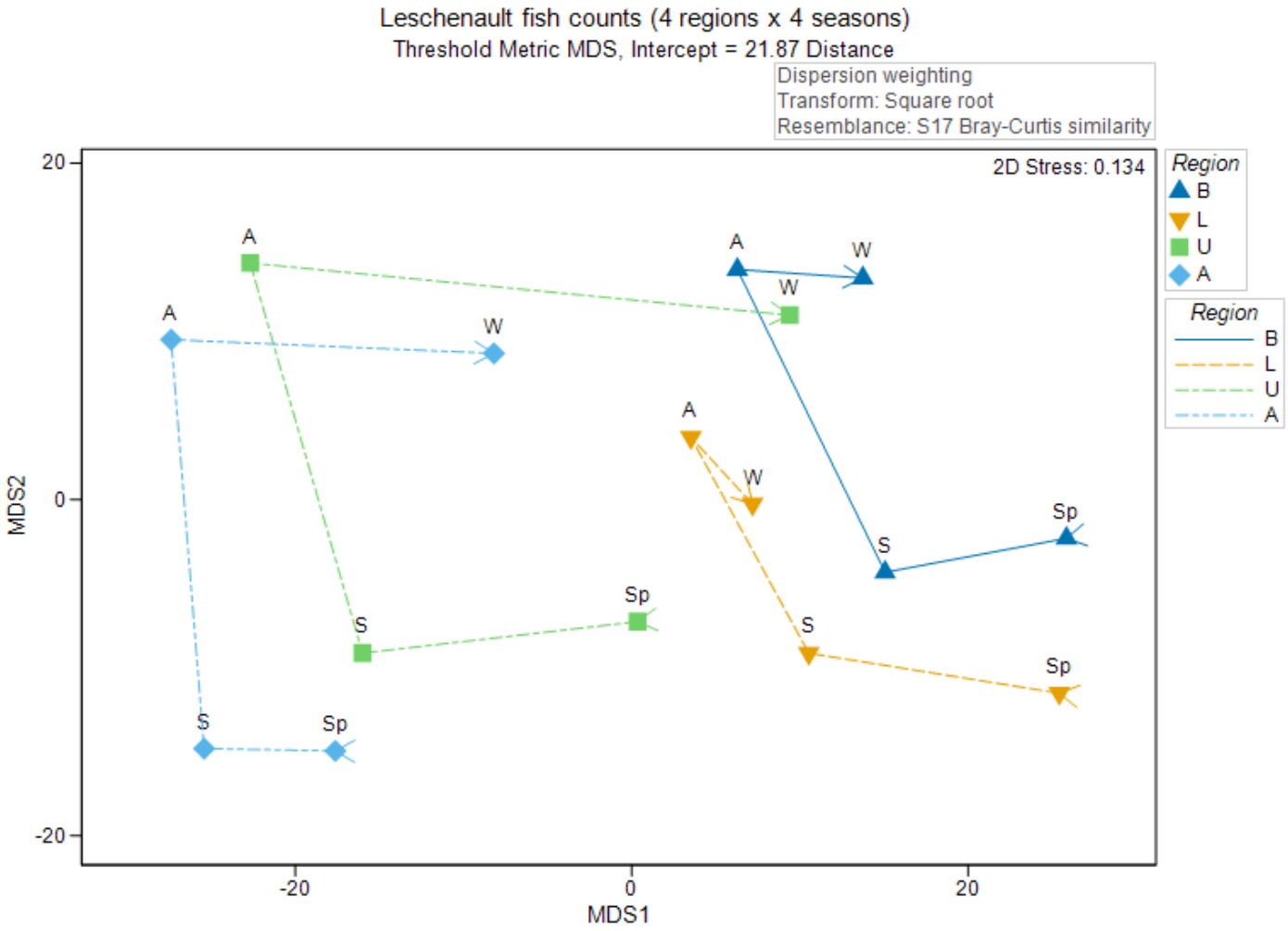


Fig. 10.8. Non-metric MDS of the Season$\times$Region centroids. Symbols correspond to the factor of 'Region' ('B' = Basal, 'L' = Lower, 'U' = Upper, 'A' = Apex), and labels correspond to the factor of 'Season' ('Sp' = Spring, 'S' = Summer, 'A' = Autumn, 'W' = Winter). A trajectory connects

the centroids sequentially through the seasons, separately within each region.

The interaction plot helpfully provides a much clearer picture of the patterns in these data by reference to the two factors. First, we can see, overall, that there is a spatial gradient of change in fish assemblages, from the Apex through to the Basal region (i.e., from left to right across the ordination plot). Second, there is a cyclical pattern of change in fish assemblages through the seasons that occurs within each of these regions (i.e., from spring, to summer, to autumn to winter).[¶]

Importantly, we can also see patterns that signal the potential reasons for detection of a significant two-way interaction here; the seasonal patterns do appear to differ for different regions of the estuary. For example, the shift from Autumn to Winter is much larger for Apex and Upper regions, compared to that observed for the Lower and Basal regions. Pair-wise comparisons confirm this observed pattern in the plot; specifically, the pair-wise test of Autumn *vs* Winter is not statistically significant for either the Basal or the Lower region (P > 0.25 in both cases), but the shift is strongly significant for the Upper and Apex regions (P < 0.001 in both cases).

Cautionary notes

1. **Measures of variability.** Just as was previously articulated for main effects plots, interaction plots of centroids should really show some measure of variability associated with each of the centroids, if possible. Some type of bootstrap or jacknife approach might be used to advantage here, but this has not yet been implemented for these plot types in PRIMER (yet). Centroids calculated from fewer samples will tend to look more variable/spread out on the plot for the reasons already discussed at the end of section 10.2. However, the sample sizes (per cell) generally do not differ much from one another for centroids shown in interaction plots, so this issue will not typically pose concerns for interpretation.
2. **Stress and interactions.** Each dataset is different, and in some cases we may not be able to easily discern the reasons behind a significant (or non-significant) interaction between two (or more) factors detected by PERMANOVA, simply by examining an interaction plot as we have done above. Although we can expect that the interaction plot will help clarify genuine patterns, our success in being able to infer finer aspects of interactions from an interaction plot will depend critically on the stress of that plot. We should always give preferential credence to the PERMANOVA results (for the main test and also for any subsequent pair-wise tests), because it operates in the space of the full dissimilarity matrix (hence has no stress).

[¶]*To do a formal test examining this pattern of cyclicity, see section 9.2 regarding tests for groups of covariates in PERMANOVA, and its application for testing cyclical models.*

10.4 Example: NZ fish assemblages

To further demonstrate the utility of main effects plots and interactions for multi-way study designs, we shall look at data from visual surveys of fish assemblages along the north-eastern coast of New Zealand completed annually (during the austral summer) over a period of 15 years, from 2001 - 2015, inclusive. In this study, divers did visual surveys at each of 4 locations ('BP' - Berghan Point, 'HP' - Home Point, 'Le' - Leigh, 'Ha' - Hahei), separated by hundreds of kilometres along the coast (Fig. 10.9). At each location, 4 sites (identified by GPS and revisited in each year) were sampled from each of two habitats: kelp forests ('k') and urchin-grazed barrens ('b') on near-shore rocky reefs. At each site, divers recorded the abundances of individual fish species observed in each of ten transects, with each transect measuring 25 m x 5 m. Over the 15-year period, a total of p = 68 different fish species were recorded during the surveys. A subset of these data (from the initial 2 years of sampling) have been analysed and discussed previously by Anderson & Millar (2004) .

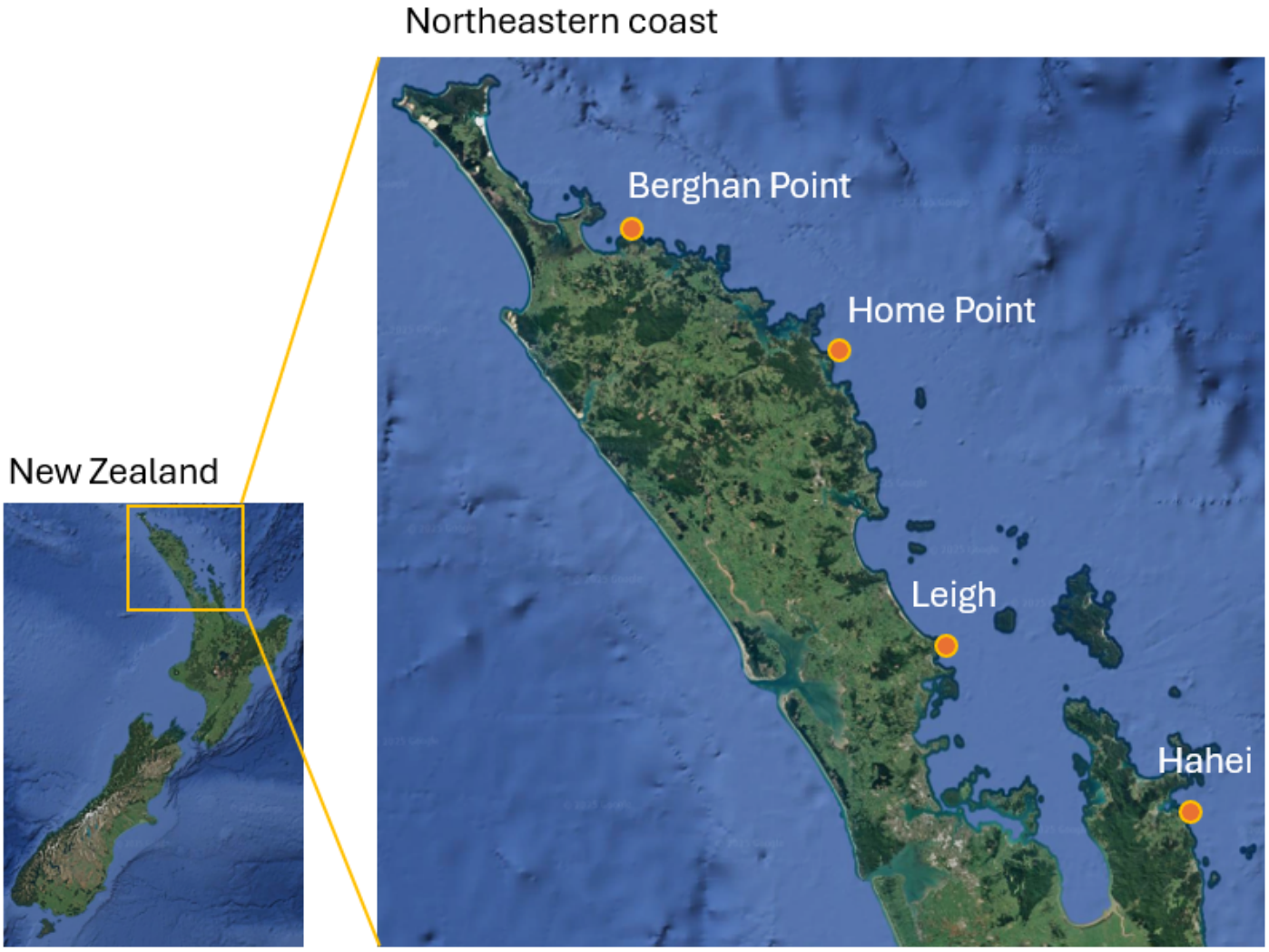
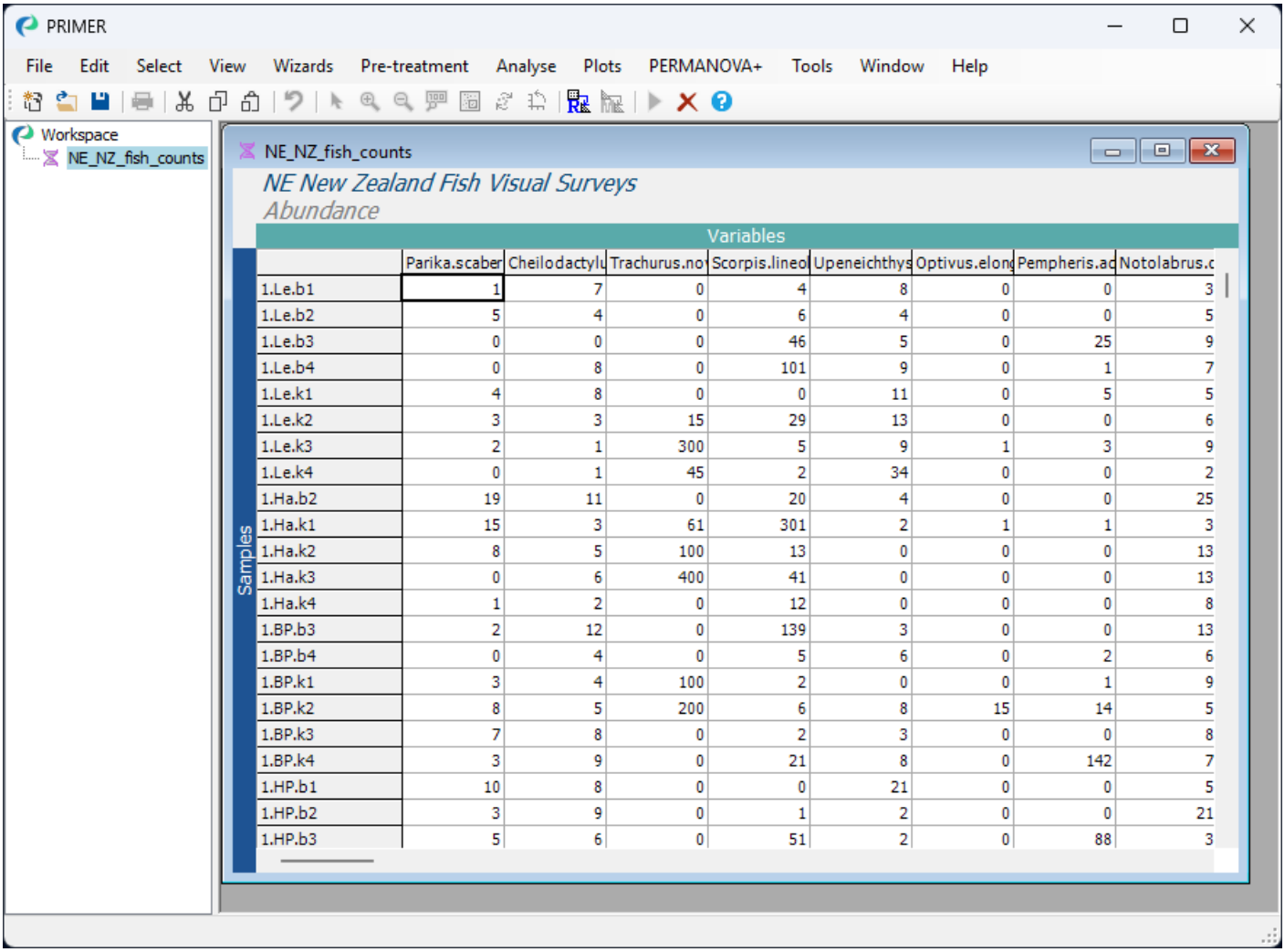


Fig. 10.9. Map of New Zealand and its north-eastern coast, showing the 4 locations where fish assemblages were surveyed annually from 2001-2015. Satellite images: Google Earth.

These data are located in the file 'NE_NZ_fish_counts.pri', found in the 'Example_P8' > 'NE_NZ_fish' folder. Note that data have been summed across the 10 transects to obtain a single measure of the multivariate fish assemblage at each site in every year.

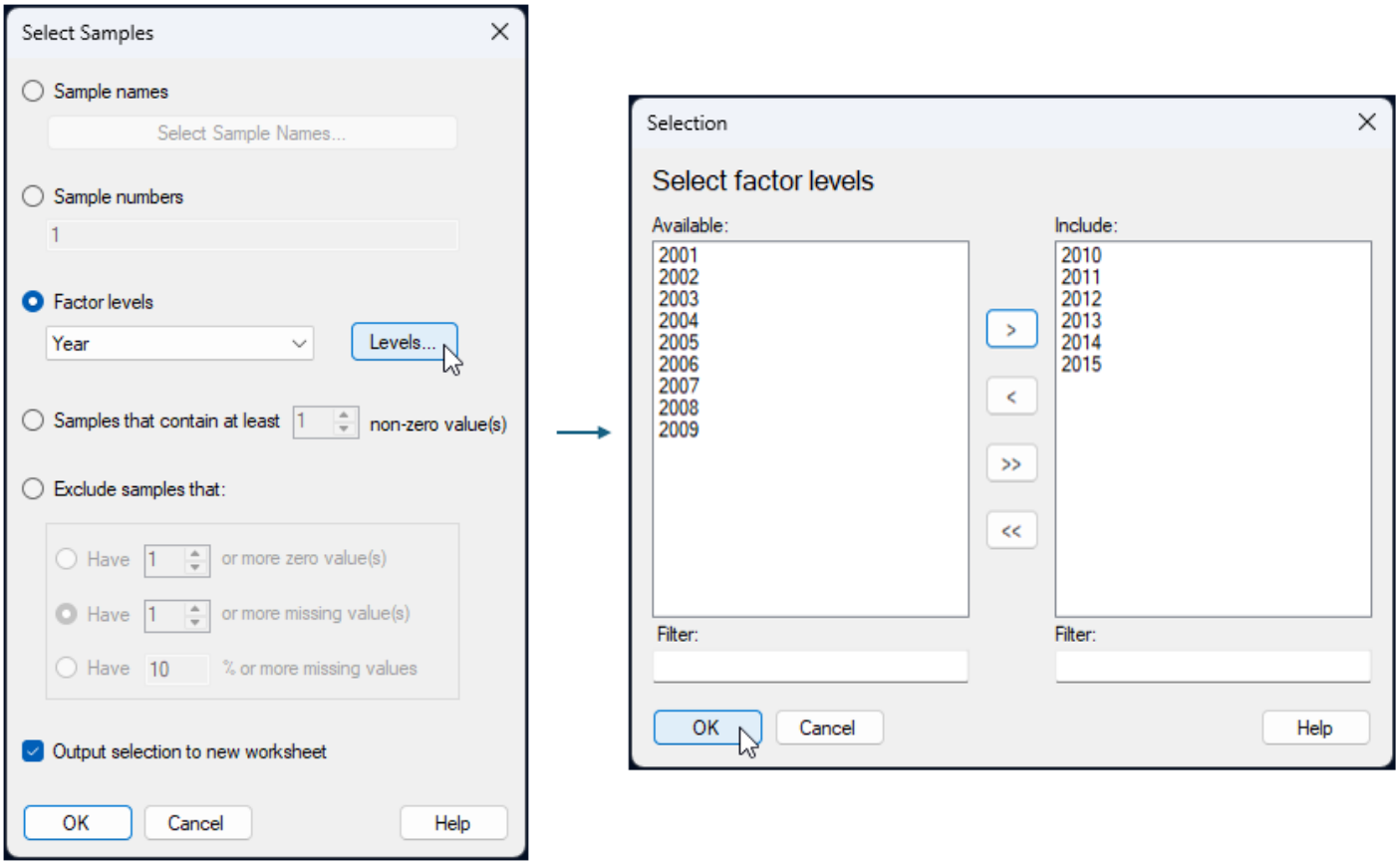
Input data, apply pre-treatments, and calculate resemblances

1. Open up PRIMER and click **File** > **Open...** to bring in the data ('NE_NZ_fish_counts', found in the 'Example_P8' > 'NE_NZ_fish' folder).



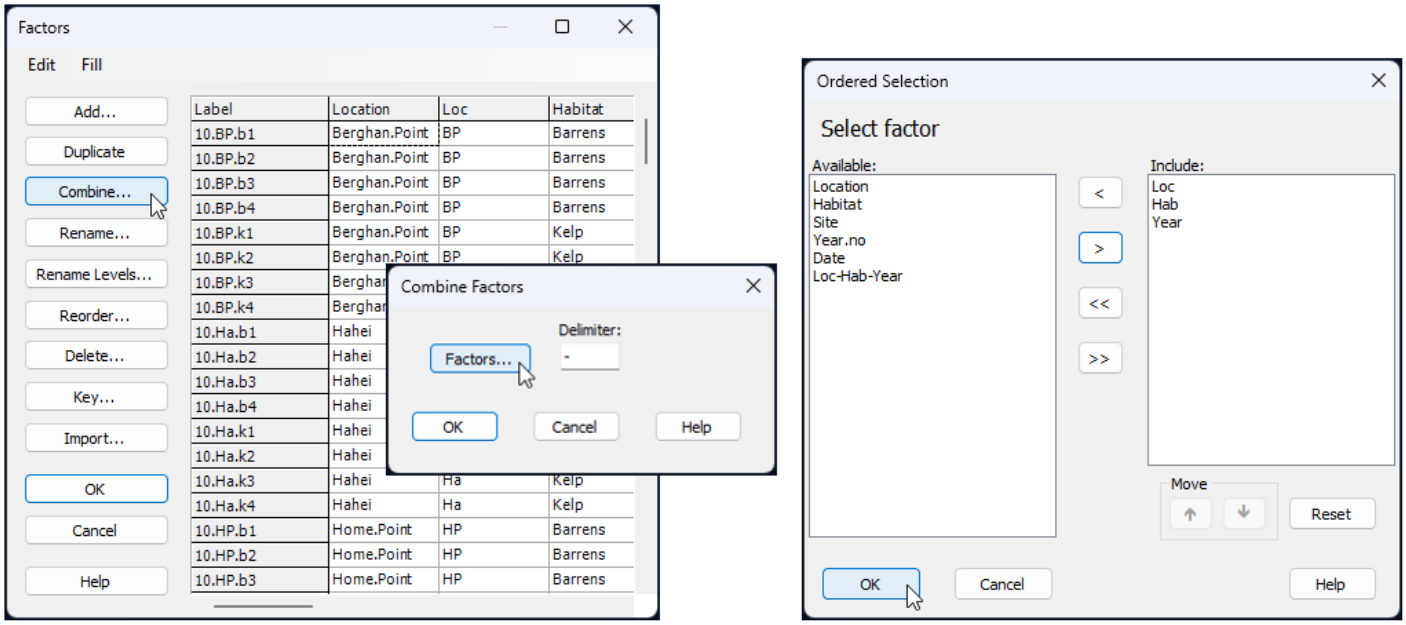
	Parika.scaber	Cheilodactylu	Trachurus.no	Scorpis.lineol	Upeneichthys	Optivus.elong	Pempheris.ac	Notolabrus.c
1.Le.b1	1	7	0	4	8	0	0	3
1.Le.b2	5	4	0	6	4	0	0	5
1.Le.b3	0	0	0	46	5	0	25	9
1.Le.b4	0	8	0	101	9	0	1	7
1.Le.k1	4	8	0	0	11	0	5	5
1.Le.k2	3	3	15	29	13	0	0	6
1.Le.k3	2	1	300	5	9	1	3	9
1.Le.k4	0	1	45	2	34	0	0	2
1.Ha.b2	19	11	0	20	4	0	0	25
1.Ha.k1	15	3	61	301	2	1	1	3
1.Ha.k2	8	5	100	13	0	0	0	13
1.Ha.k3	0	6	400	41	0	0	0	13
1.Ha.k4	1	2	0	12	0	0	0	8
1.BP.b3	2	12	0	139	3	0	0	13
1.BP.b4	0	4	0	5	6	0	2	6
1.BP.k1	3	4	100	2	0	0	1	9
1.BP.k2	8	5	200	6	8	15	14	5
1.BP.k3	7	8	0	2	3	0	0	8
1.BP.k4	3	9	0	21	8	0	142	7
1.HP.b1	10	8	0	0	21	0	0	5
1.HP.b2	3	9	0	1	2	0	0	21
1.HP.b3	5	6	0	51	2	0	88	3

2. To simplify matters a little, we will analyse a subset of the data here. Select only years 2010 - 2015 inclusive. From the 'NE_NZ_fish_counts' sheet in PRIMER, click **Select** > **Samples...**, tick the box to '$\checkmark$ Output selection to new worksheet', then choose $\bullet$ Factor levels 'Year', and click the 'Levels' button (Levels...). In the 'Selection' dialog, click on each of the years from 2010 - 2015, then click on the right arrow (>) to move them over into the 'Include' box on the right, then click '**OK**'. These steps are shown below:



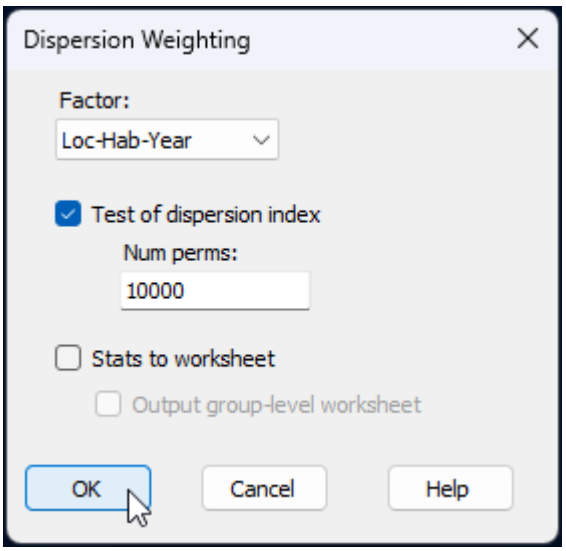
The subsetted data will now be found in the sheet named 'Data1' in the Explorer tree. We shall use dispersion weighting (as many fish species naturally occur in aggregations/schools; see Clarke *et al.* (2006a)), followed by a square-root transformation, to pre-treat the raw data prior to analysis. For the dispersion weighting, we need to provide several groups of replicates. A natural grouping factor to use here is the one we can create as a combination of the 3 main factors in this study design: Location, Habitat and Year. We first have to create this combined factor.

3. From the 'Data1' sheet, click **Edit** > **Factors...**, then click on the 'Combine' button (Combine...), followed by the 'Factors...' button (Factors...). In the 'Ordered Selection' dialog, click on each of the factors: 'Loc', 'Hab' and 'Year', in turn, then click on the right arrow (>) to move them over into the 'Include' box on the right, then click '**OK**'(see below):



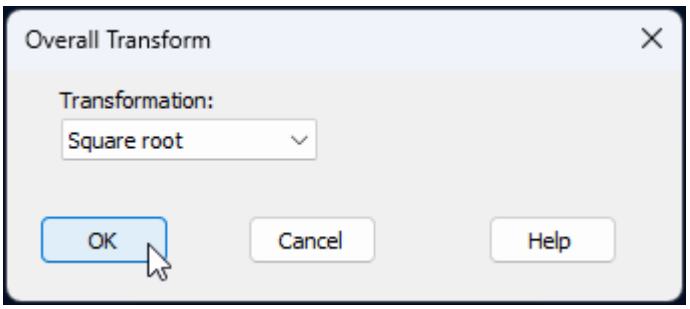
The resulting combined factor is called 'Loc-Hab-Year' and can be seen associated with the 'Data1' sheet when we click **Edit** > **Factors...** (it is in the last column).

4. Now we can apply dispersion weighting on the basis of this combined factor. From the ' Data1' data sheet, click **Pre-treatment** > **Dispersion Weighting...** and in the resulting dialog, choose Factor: Loc-Hab-Year, then click '**OK**'.



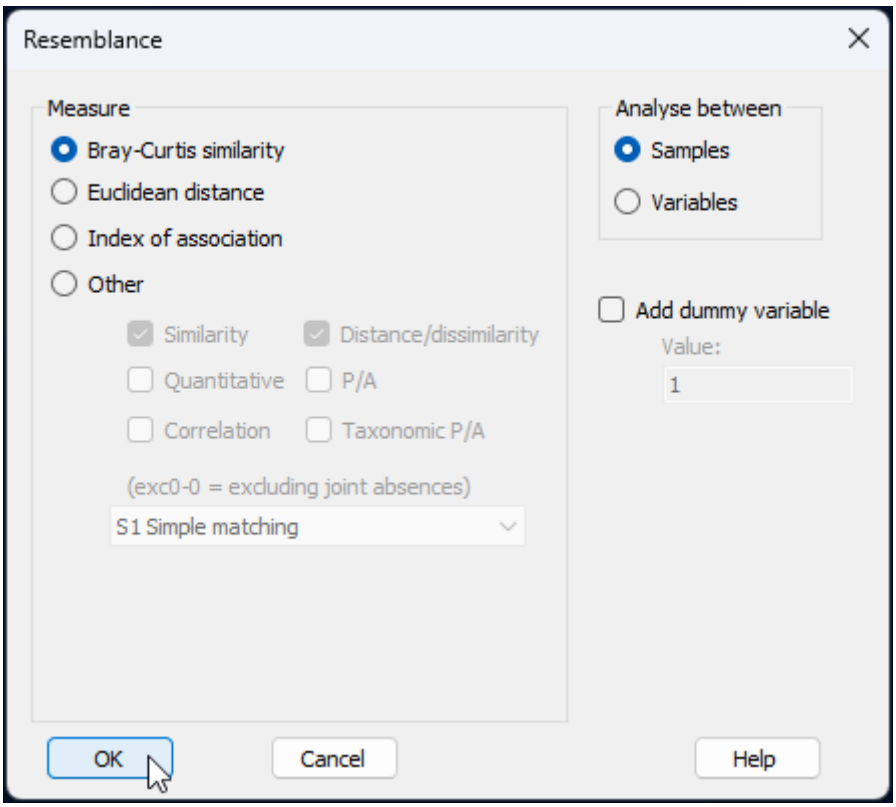
The dispersion-weighted data are provided in the sheet named 'Data2' in the Explorer tree.¶

5. Next, apply a square-root transformation. From the 'Data2' sheet, click **Pre-treatment** > **Transform(overall)...** and choose Transformation: Square root, then click '**OK**'.



The pre-treated data (square-root transformed dispersion-weighted values) are provided in the sheet named 'Data3' in the Explorer tree. We are ready now to calculate resemblances and proceed with analyses from there.

6. Calculate Bray-Curtis resemblances among all sampling units using the Bray-Curtis measure. From the 'Data3' sheet, click **Analyse** > **Resemblance...**, take all of the defaults, and click '**OK**'.



The resulting resemblance matrix is in the item called 'Resem1' in the Explorer tree.

PERMANOVA

We shall begin by analysing these data using PERMANOVA on the basis of the Bray-Curtis resemblance matrix calculated from square-root transformed dispersion-weighted data ('Resem1'). Doing PERMANOVA first, in cases where there is a multi-factorial design, *before* we embark on ordinations, is often a good idea. The PERMANOVA partitioning will give us a quantitative comparative analysis of the relative importance of the factors (and their interactions) in our study design. Knowing which factors 'matter' (and which ones matter 'most') will help point us towards useful ordinations, to focus on the primary structuring factors of interest. Specifically, seeing the PERMANOVA output can help us decide which factors are worth looking at in more detail with main effects and interaction plots.

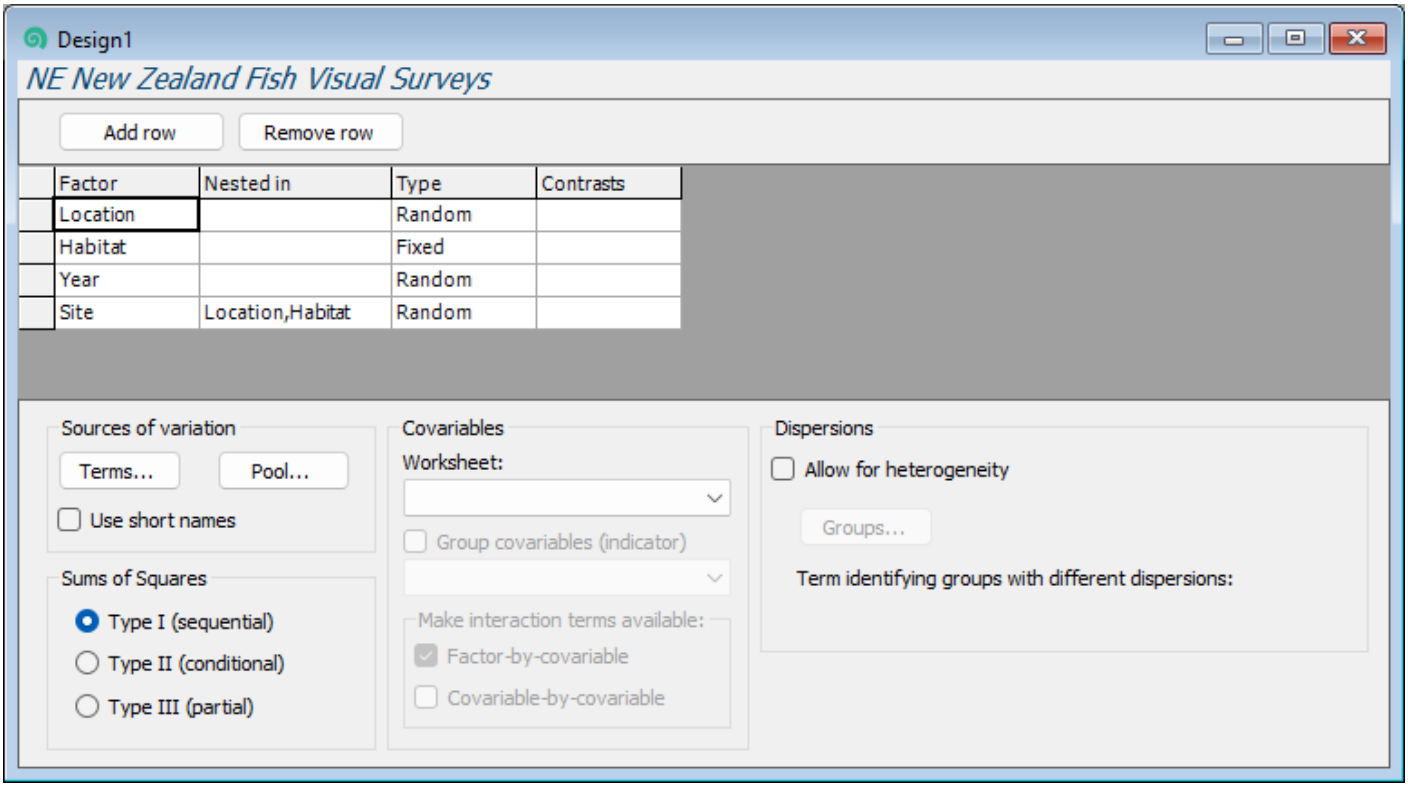
Create the PERMANOVA design file

There are four factors in this PERMANOVA design, as follows:

- **Location** (random with 4 levels: Berghan Point, Home Point, Leigh and Hahei)

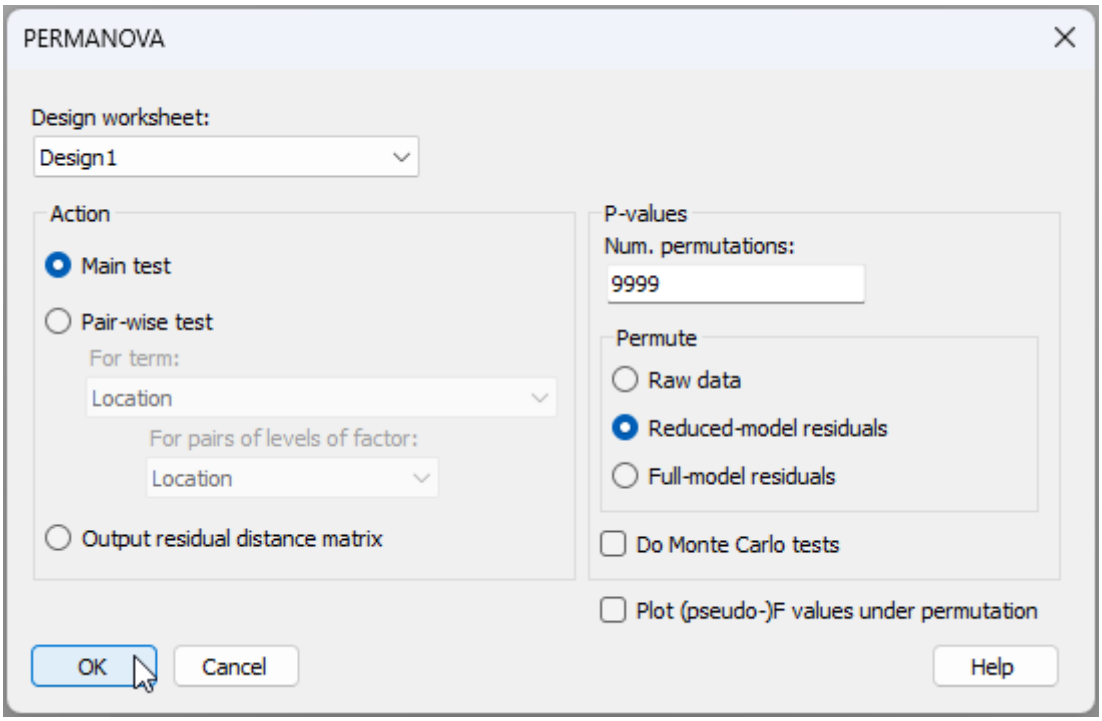
- **Habitat** (fixed with 2 levels: kelp forest and urchin-grazed barrens)
- **Year** (random with 15 levels, a subset of 6 levels are examined here: 2010 - 2015 inclusive)
- **Site** (random and nested in Location and Habitat)

7. From the 'Resem1' matrix, click **PERMANOVA+** > **Create PERMANOVA Design...**. In the resulting design file (called 'Design1'), click the 'Add row' button (Add row) three times, so there will be four rows in total here, then double-click inside each cell in the first column and choose the following factors in the study design ('Location', 'Habitat', 'Year', and 'Site'), in turn. Specify the 'Type' and 'Nested in' structure, as per the above 4-factor design by double-clicking inside the relevant cell in those columns for each factor in turn. The resulting design file, which mirrors the above articulated design, should look like this:



Run the PERMANOVA model

8. From the 'Resem1' matrix, click **PERMANOVA+** > **PERMANOVA...**, make sure the 'Design worksheet:' is 'Design1', go with all of the default options for the rest, and click '**OK**'.‡



The resulting PERMANOVA output file ('PERMANOVA1') will look like this:

PERMANOVA1

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
Location	3	46803	15601	5.0144	0.0001	9803
Habitat	1	27422	27422	5.7783	0.0016	9947
Year	5	15421	3084.2	2.352	0.0001	9873
LocationxHabitat	3	10432	3477.3	1.4495	0.0035	9811
LocationxYear	15	19656	1310.4	1.8105	0.0001	9746
HabitatxYear	5	7149.3	1429.9	1.497	0.0263	9896
Site(LocationxHabitat)	24	46582	1940.9	2.6856	0.0001	9685
LocationxHabitatxYear	15	14294	952.92	1.3188	0.0047	9717
Res	119	85985	722.56			
Total	190	2.7374E+05				

Estimates of components of variation

Source	Estimate	Sq.root
V(Location)	273.71	16.544
S(Habitat)	245.77	15.677
V(Year)	55.712	7.4641
V(LocationxHabitat)	54.606	7.3896
V(LocationxYear)	73.74	8.5872
V(HabitatxYear)	29.893	5.4675
V(Site(LocationxHabitat))	204.32	14.294
V(LocationxHabitatxYear)	58.021	7.6171
V(Res)	722.56	26.88

In a nutshell, we see here that all of the factors are statistically significant in this model. There is a significant three-way 'Location$\times$Habitat$\times$Year' interaction ($F_{15,119}$ = 1.32, P < 0.01). This suggests that an interaction plot involving those three factors would be useful to examine.

Note that the 'Sq.root' column in the section labelled '*Estimates of components of variation*' of the PERMANOVA output file gives us sizes of effects (in Bray-Curtis units) for each term in the model. For the main effects, this can be interpreted as a measure of the average (or 'standard') deviation

of any given group centroid for that factor from the overall centroid.

Over and above the large variation among Sites (having a square-root estimated component of variation > 14 Bray-Curtis units), the main effects of Location and Habitat are also particularly strong (with square-root components of 16.5 and 15.7 Bray-Curtis units, respectively), while the main effects of Year were quite a lot smaller in size (7.5 Bray-Curtis units). This suggests that it would also be useful to examine a main effects plot involving those three factors (Location, Habitat and Year) for additional insights.

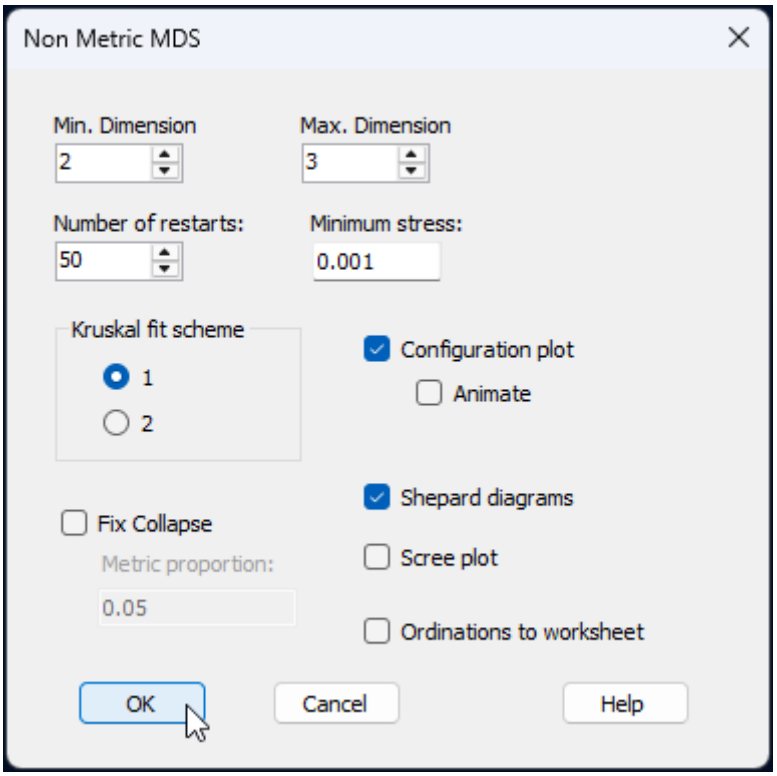
Ordinations

To help us visualise relationships among the sampling units, and potential structuring due to the spatial and temporal factors in the study design, we might consider plotting:

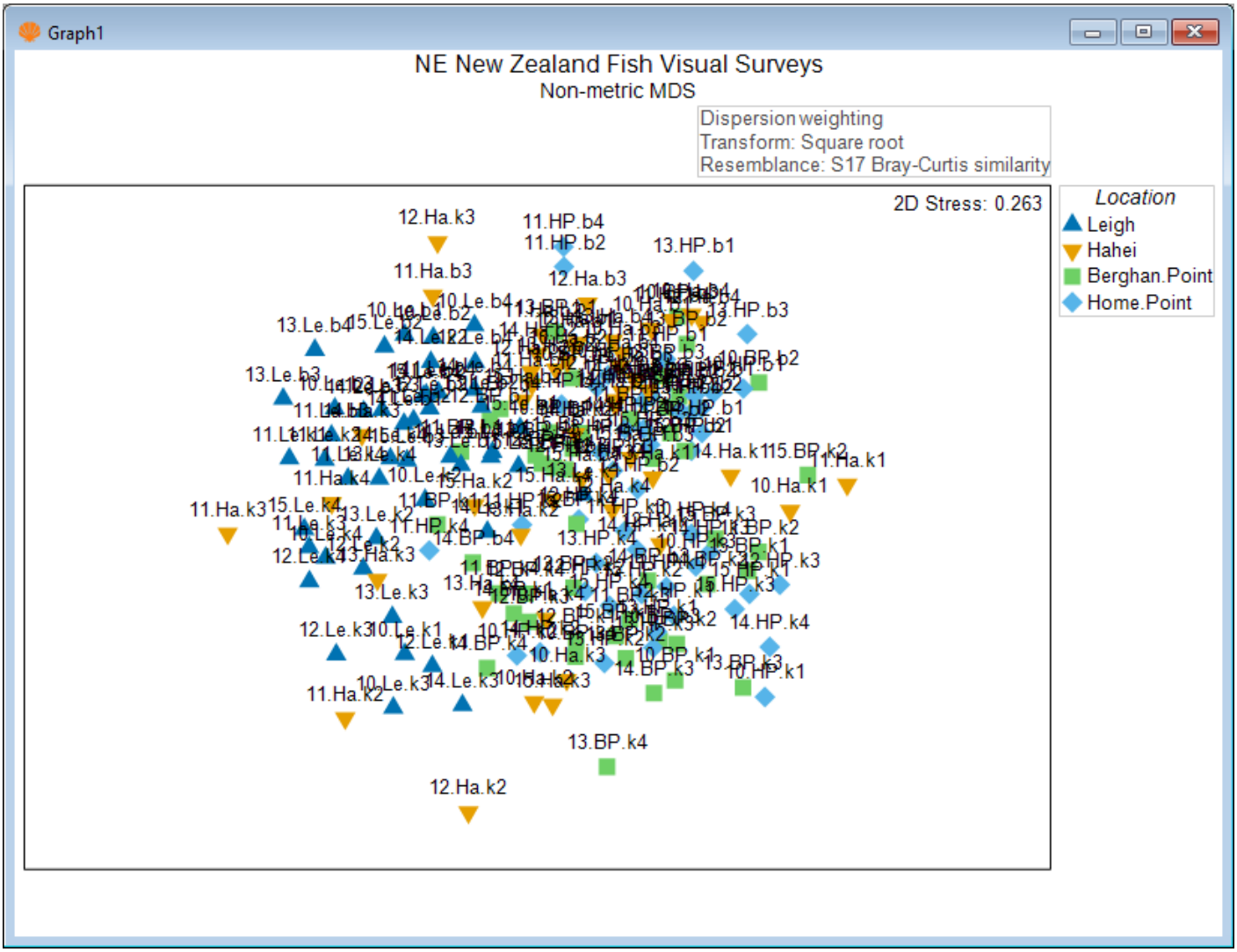
- an ordination of *all sampling units* from the full resemblance matrix;
- an ordination of centroids for the main factors: Location, Habitat and Year (a *main effects plot*);
- an ordination of centroids corresponding to combinations of factor levels for Location-by-Habitat-by-Year (an *interaction plot*).

Ordination of all sampling units

9. From the 'Resem1' matrix, click **Analyse** > **MDS** > **Non-metric MDS (nMDS)...**, take all of the default options and click '**OK**'.



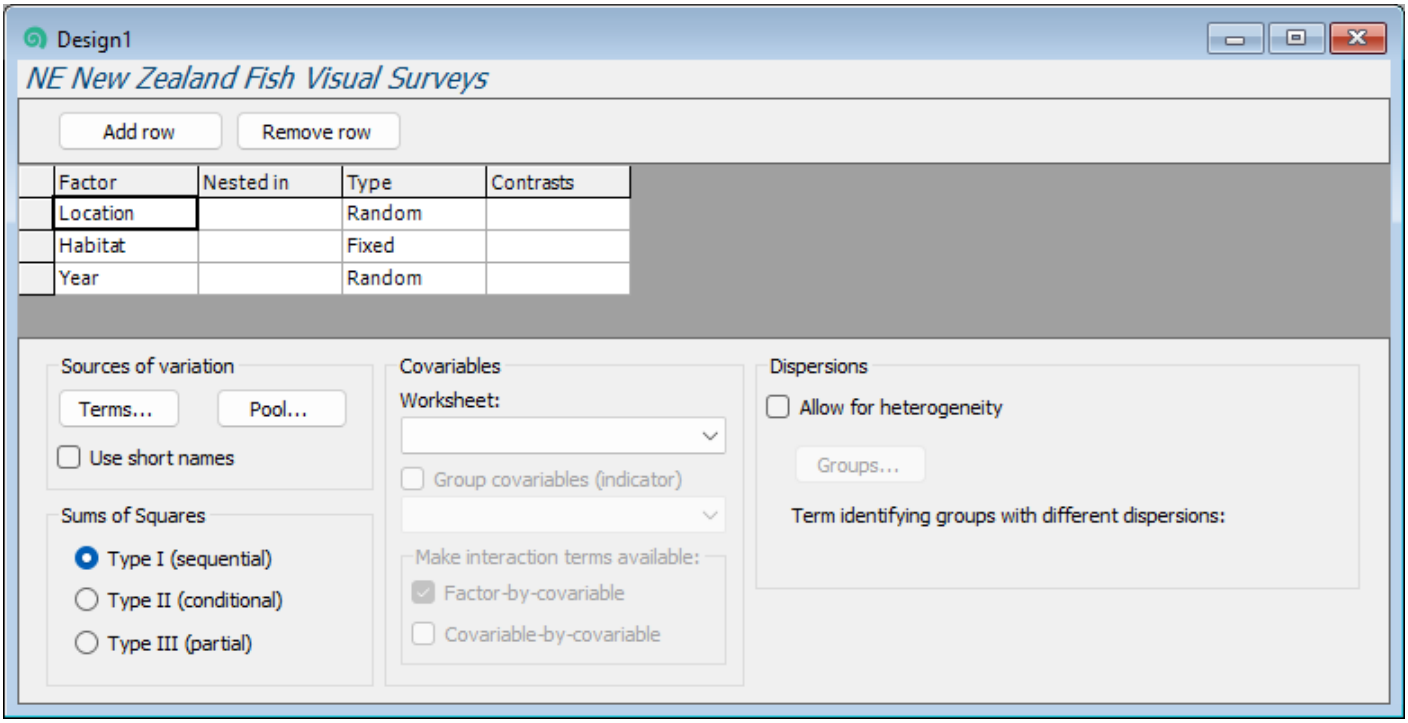
The resulting 2D nMDS ordination (provided in 'Graph1') looks like this (by default):



This is exceptionally messy and uninterpretable. Even if we remove the labels, no useful information can be taken from this ordination; for a start, the stress is just way too high (> 0.26). Even the best 3D nMDS solution has quite high stress (0.197) and looks like little more than a 'ball of fuzz'.

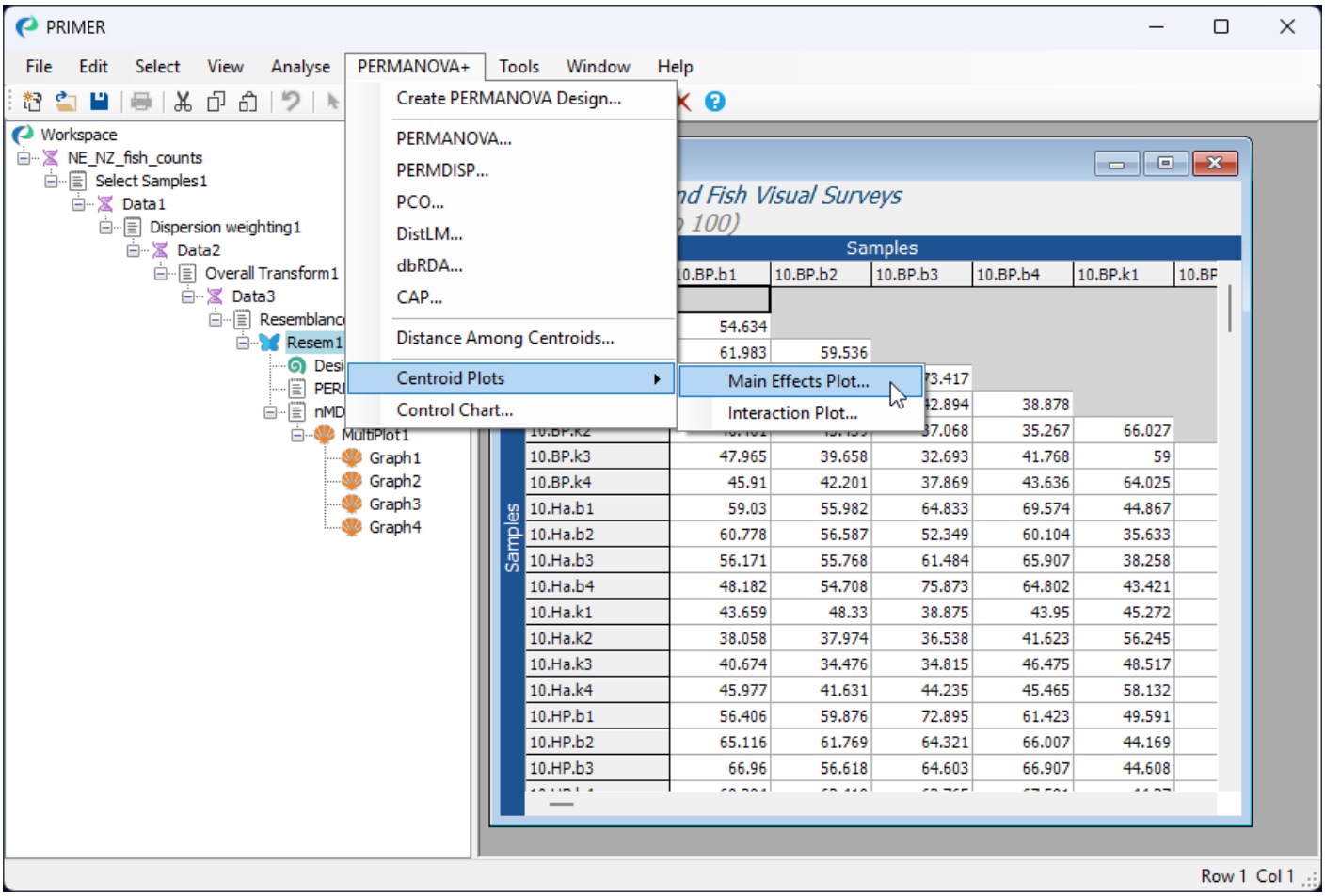
Main effects plot

10. To obtain a main effects plot, we first have to ammend our design file to remove the 'Site' factor, so as to focus our attention on the three main effects of interest here: 'Location', 'Habitat' and 'Year'. For this, simply click once on row 4 to highlight it, then click the 'Remove row' button (Remove row) to remove the fourth (and final) row. The resulting 3-row design file (still called 'Design1') should look like this:[§]

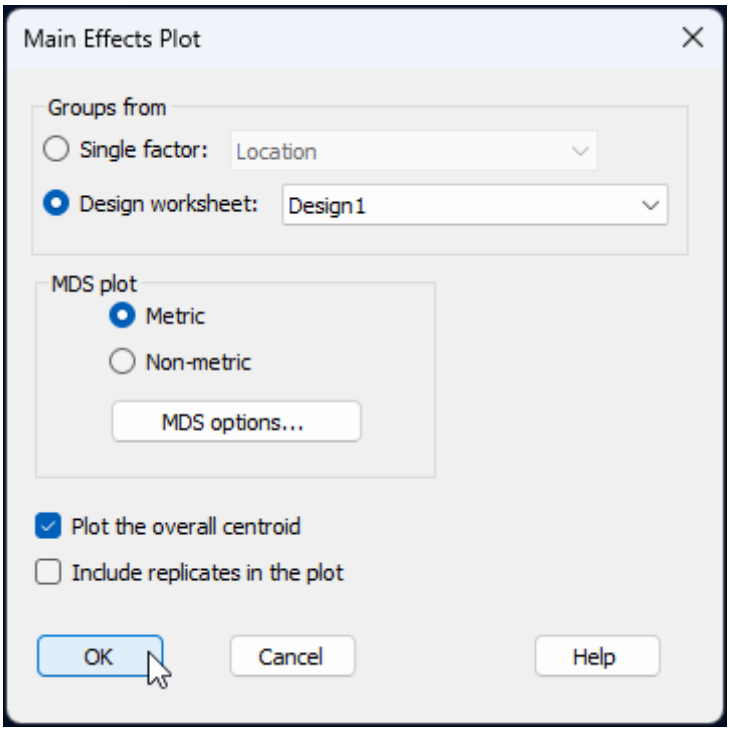


As an aside, note that the 'Type' of each factor has no particular relevance or impact on a main effects plot.

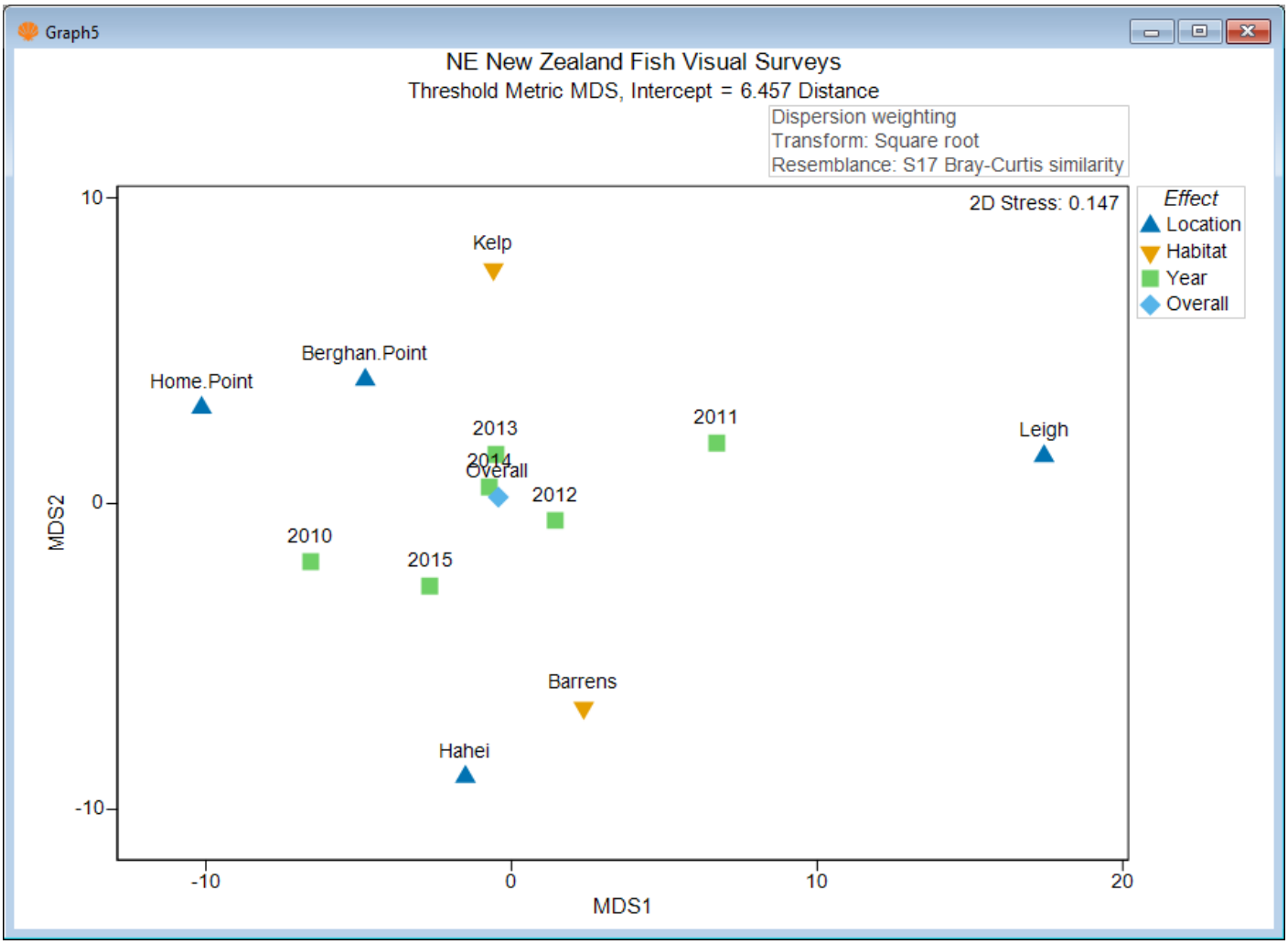
11. From the 'Resem1' matrix, click **PERMANOVA+** > **Centroid Plots** > **Main Effects Plot...**.



Choose (Groups from $\bullet$Design worksheet: Design1) and (MDS plot $\bullet$Metric). Also choose to $\checkmark$Plot the overall centroid, then click '**OK**', like so:



The resulting main effects plot for these three factors is a threshold-metric MDS (called 'Graph5' in the Explorer tree), as shown below.

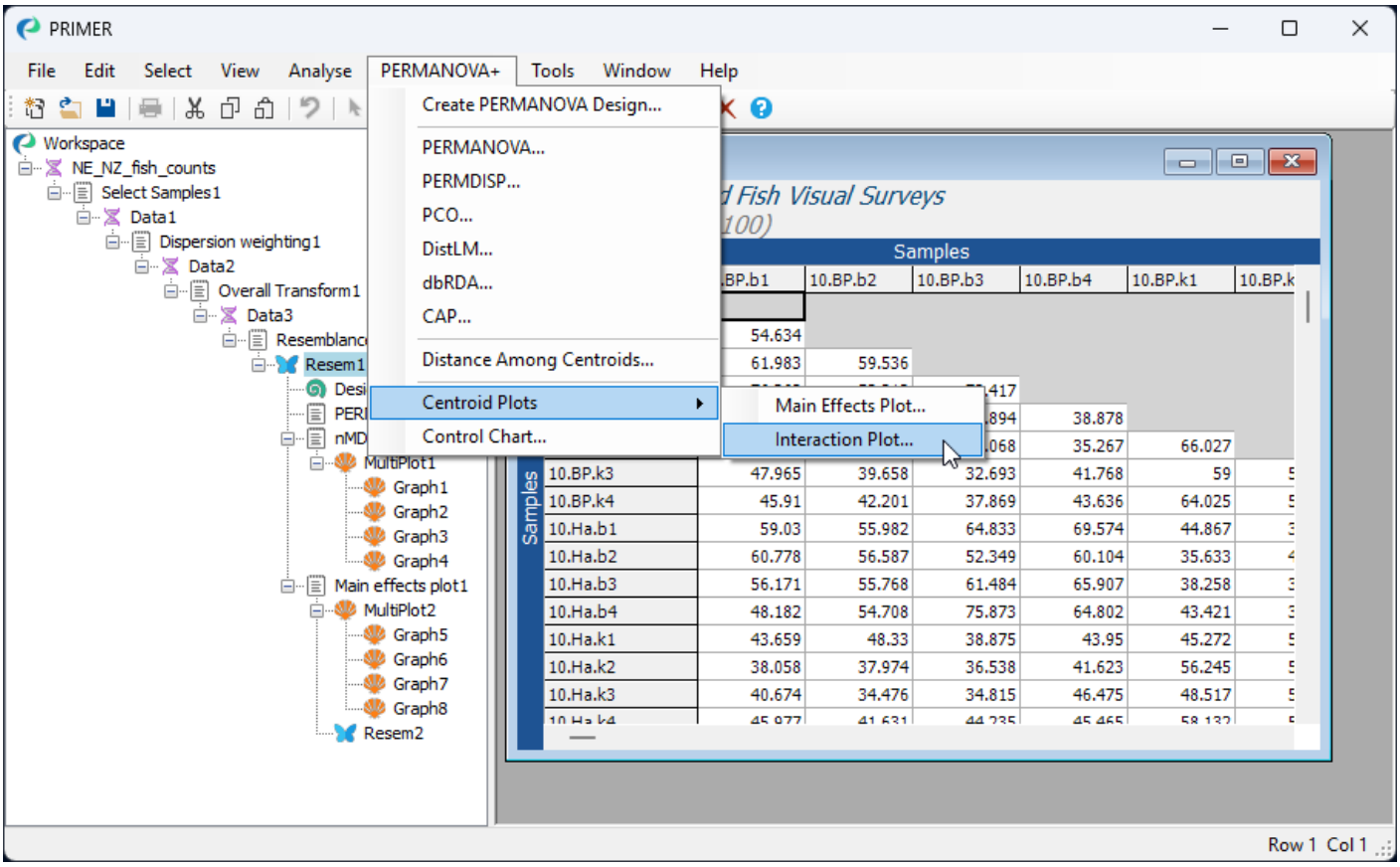


Outlined below are a number of observations we can draw from this ordination of main-effect centroids.

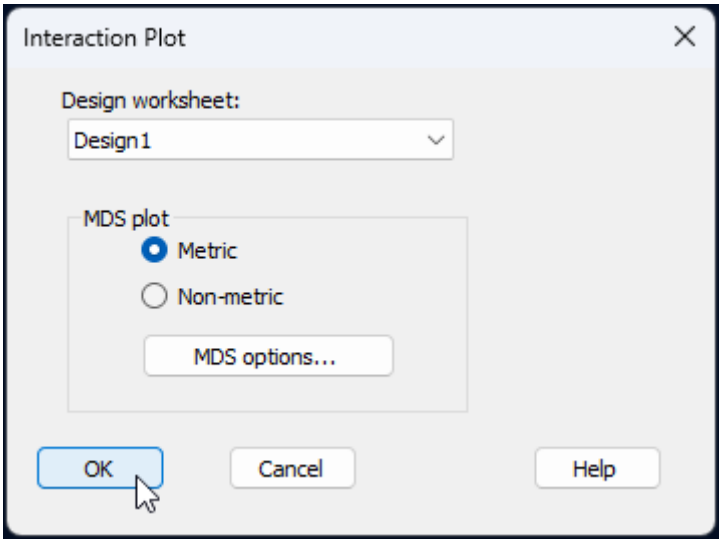
- First, the effects due to different locations are rather large, compared to the other factors. Fish assemblages from Leigh appear to be quite different from those at other locations, while fish assemblages from Berghan Point and Home Point (the two northern locations) are more similar to one another.
- Second, there is also quite a strong effect of Habitat. The effects of habitat on these fish assemblages (i.e., the distances from each of the habitat centroids to the overall centroid) appear to be similar in size to the location effects, on average, although they occur in a different direction in the multivariate space.
- Third, the inter-annual (temporal) effects are less important (smaller in size) than the spatial effects due to either the location or the habitat factors.
- Finally, and more generally, the lengths of the deviations of individual centroids for different levels of a given factor from the overall centroid, as may be estimated in this tmMDS ordination, are similar in size (and match the rank order of relative sizes) for these three main effects, as given in the PERMANOVA output file. That is, Location (~15-17 units) > Habitat (~13-15 units) > Year (~6-10 units). (Don't forget to add on the intercept value from the tmMDS to these in order to get the full picture regarding these distances). Of course, we mustn't forget that there is *stress* in this ordination diagram, so mean distances to the overall centroid won't match the effect sizes given by PERMANOVA precisely (we trust the quantitative information given by PERMANOVA more, where there is no stress involved), but for purposes of visualising the relative importance of these factors, this is really, nevertheless, a very helpful plot.

Interaction plot

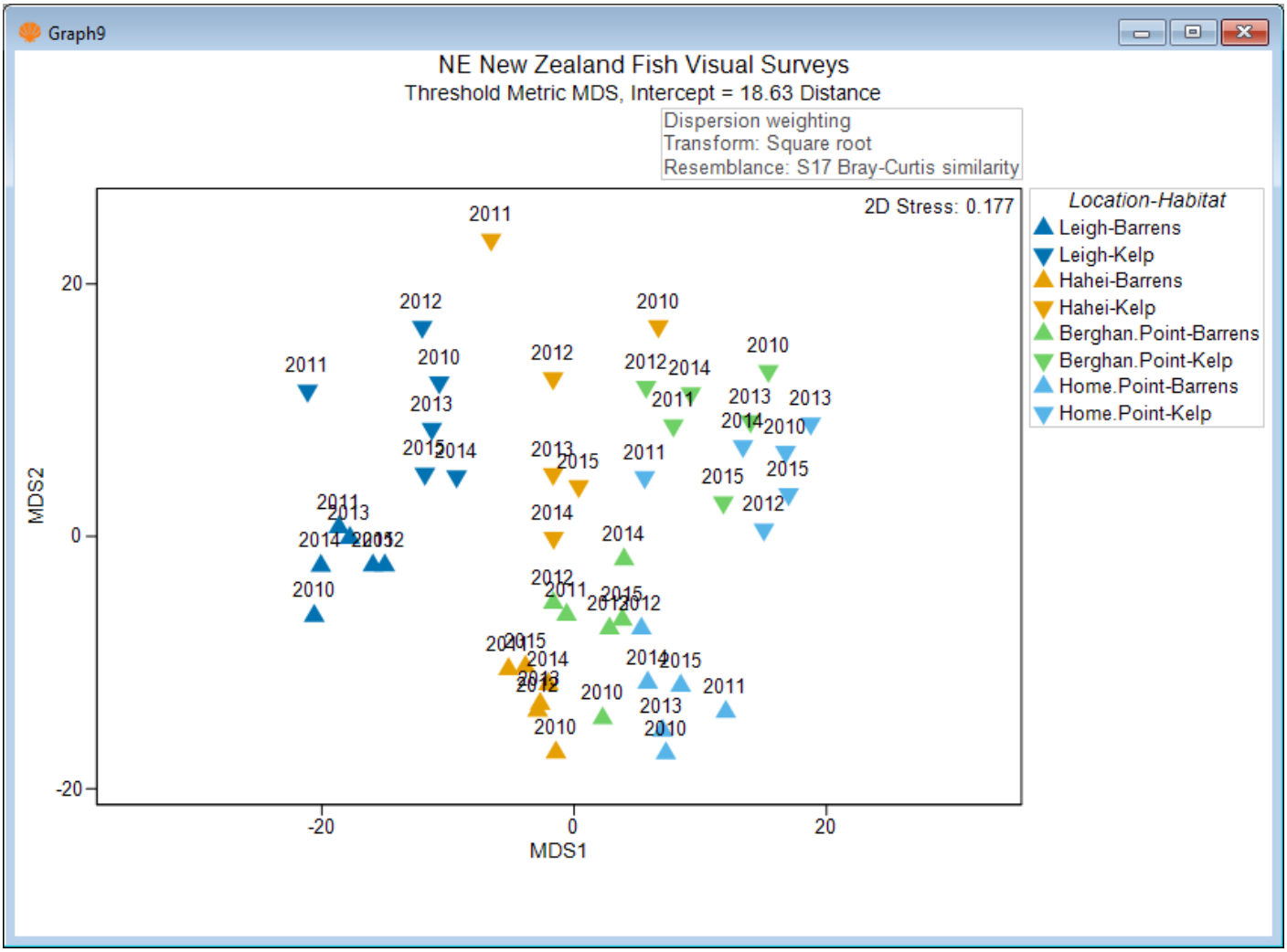
12. To obtain a 3-way interaction plot, go to the 'Resem1' resemblance matrix in the Explorer tree, and click **PERMANOVA+** > **Centroid Plots** > **Interaction Plot...**.



In the resulting dialog, choose (Design worksheet: 'Design1') and (MDS plot $\bullet$Metric), then click '**OK**'.



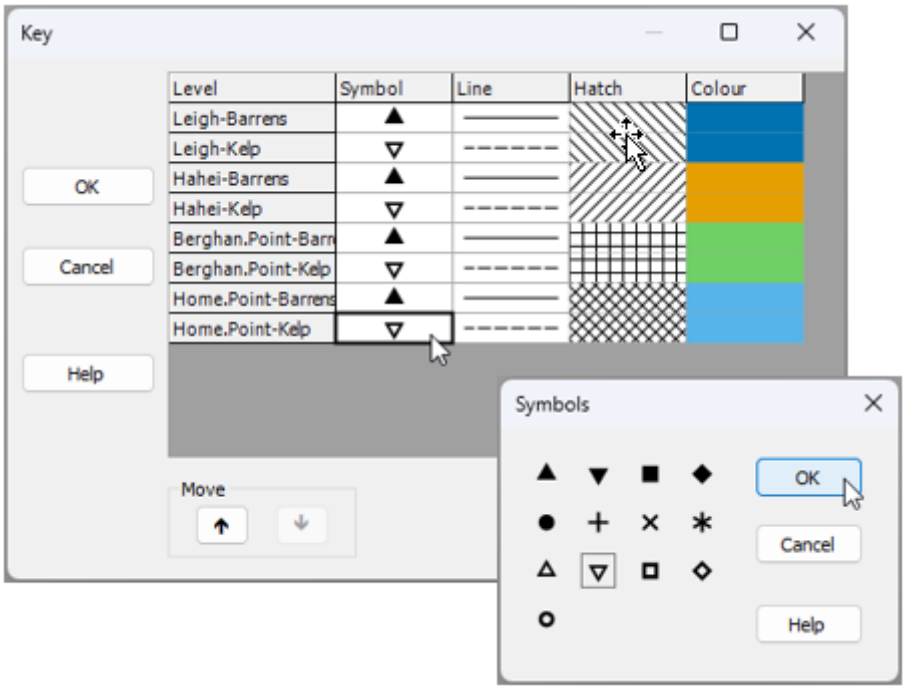
This will produce a three-way interaction plot ('Graph9') that looks like this:



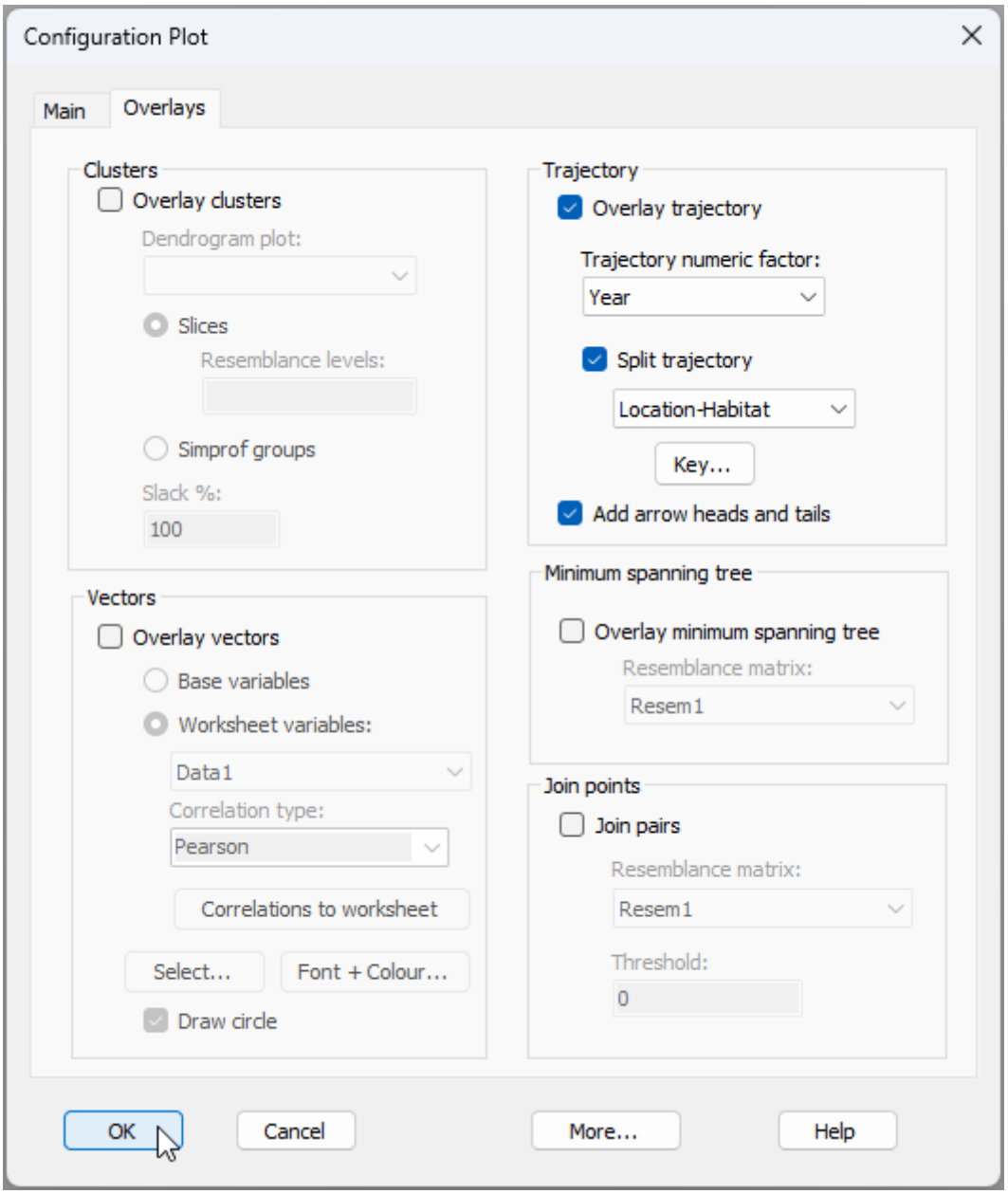
Note that what we see in this plot will depend on the order of the factors in the Design file. The default actions implemented by the Interaction plot routine are as follows:

- The **first** factor listed in the design file will be used to specify the **colours** of symbols.
- The **second** factor listed in the design file will be used to specify the **shapes** of symbols (i.e., varying across each colour).
- Thus, **symbols** are created as a *combination* of the first 2 listed factors in the design file.
- The **third** factor listed in the design file will be used to specify the **labels**.

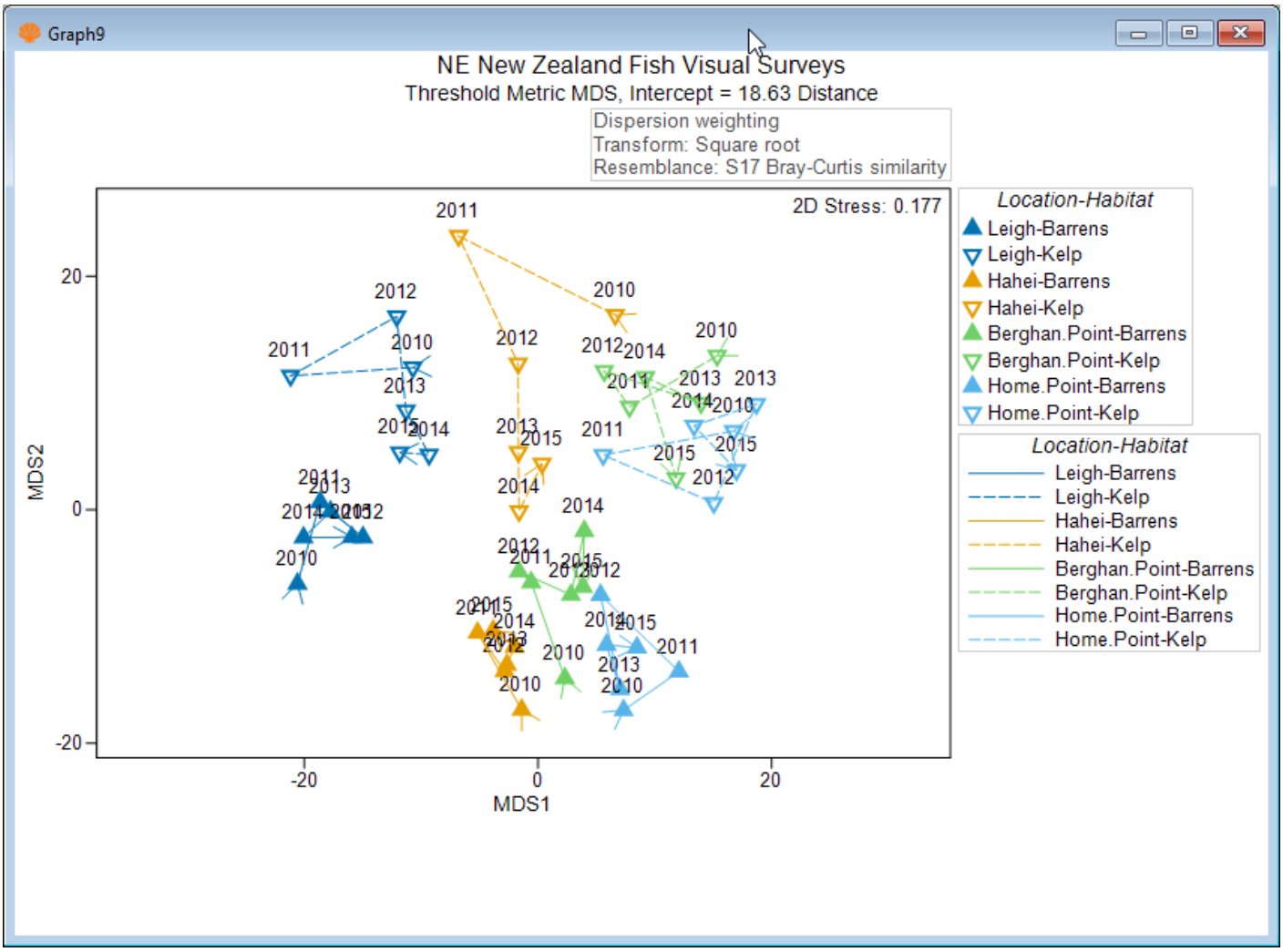
In this particular example, the 'Habitat' factor is a little difficult to discern *via* the default chosen symbol shapes (i.e., upward triangle *vs* downward triangle). As there are only 2 levels of the 'Habitat' factor, we could, instead, modify the default output by make the symbols corresponding to 'kelp' habitat open symbols instead of closed symbols (leaving the 'barren' habitat symbols as they are). This is easily done by clicking on the legend inside the graphic, and using the resulting dialog to modify the relevant symbols, like so:



Another improvement to this graphic would be to superimpose a set of trajectories that link consecutive years for each Location-by-Habitat combination. Click **Graph** > **Special..**, and under the '**Overlays**' tab, tick the box to $\checkmark$Overlay trajectory, with Trajectory numeric factor: Year and $\checkmark$Split trajectory by Location-Habitat, then click **OK**, as shown below:



The improved version of the interaction plot now looks like this:



This ordination is clearly far more useful than the nMDS ordination of all sampling units. The patterns we can see here include (but may not be limited to) the following:

- Fish assemblages at Leigh (dark blue) are quite distinct from those found at the other locations sampled along the coast, while the fish assemblages at the two northern locations (Berghan Point in light blue, and Home Point in green) are more similar to one another.
- There are clear differences in fish assemblages found in kelp forests *vs* those in barrens habitats (open symbols *vs* closed symbols, respectively), and these appear to be broadly similar effects across all 4 of the locations.
- Variation among years (e.g., look at the lengths of the trajectories) appears to differ for different locations and habitats. For example, temporal variation among years appears to be somewhat larger (years are more dispersed) in kelp habitats *vs* in urchin-grazed barrens habitats, and this difference is more marked for Leigh and Hahei, compared to Home Point and Berghan Point. Such patterns may easily explain the significant three-way interaction detected by PERMANOVA.

¶*A quick glance at the results file from this operation (called 'Dispersion weighting1' in the Explorer tree), shows how useful this pre-treatment operation was. A large number of these fish species have divisors that are a lot larger than 1, indicating they are highly aggregated. Many of these fish are indeed schooling fish, and some have numbers of individuals that are typically in the hundreds*

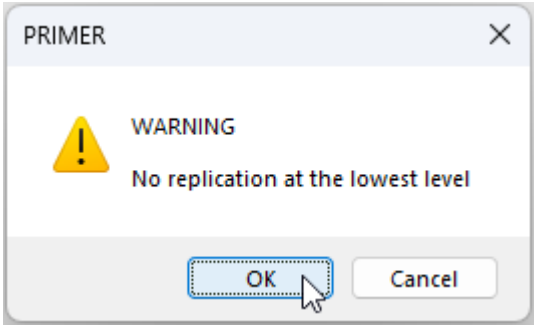
or even thousands within a single school. Recall that dispersion weighting downweights large abundance values for variables that have high variance-mean ratios, i.e., that are more erratic in their statistical behaviour. See Clarke et al. (2006a) for more details.

---

†*It is not strictly necessary to perform any kind of statistical analysis (whether it be PERMANOVA or ANOSIM, etc.) prior to creating ordination plots, of course. It is just sometimes helpful when there are a lot of factors. We can use the results from the PERMANOVA analysis to inform us of the most salient structuring factors in the study design and whether they interact, thus, pointing us towards what might be some of the most appropriate and useful ordinations to look at.*

---

‡ *For this design, you will see a warning sign pop up, to highlight the fact the there is no replication at the lowest level. We know that there were, in fact, 10 replicate transects per site, but recall that we have summed these up to the site level for the analysis of fish communities.*



The pop-up warning means we cannot estimate the 'Year$\times$Site(Location$\times$Habitat)' interaction term because it is indistinguishable from the residual variation. You can click 'OK' in response to this warning, and PERMANOVA will continue with the analysis, automatically removing this inestimable high-order interaction from the design. It is worth bearing in mind that what is called 'Residual' in the resulting PERMANOVA output file is actually a mixture of the 'Residual' variation (from one site to another within a given year) and the potential variation due to the 'Year$\times$Site(Location$\times$Habitat)' interaction. This mixture is actually perfectly fine for generating all of the relevant tests we need for other terms of interest in our study design, so it is not something that will cause us to lose any sleep. All terms in our model are tested with perfect statistical rigour. As a further aside, note that 'Site' could optionally have been set up as a factor of type 'Subject/Whole-plot error', because we repeatedly went back to sample the same sites through time (every year), just as in a repeated-measures design (see section 8.3). If we had indeed chosen for 'Site' the 'Type' of 'Subject/Whole-plot error' in our design file, then no such warning would appear. However, we preferred here to include 'Site' as a nested factor, for transparency in the assumptions attending our model.

---

§*By the way, it is ok to keep all four factors in the design when you run the main effects plot, if you like. If you do that, there will be a centroid for every unique site in the resulting plot as well, so the ordination will just be a bit more busy to look at and make sense of. Keep in mind, if you do that, however, that the 'Site' centroids deviate around the 'Location-by-Habitat' centroids in the PERMANOVA model, and without the latter being plotted explicitly, site-level effects will be more difficult to read directly from the plot.*