18. MULTIDIMENSIONAL SCALING AND CLUSTER ANALYSIS

In the previous chapter, we were mainly interested in *R*-mode analyses that were based on associations between variables and scaled objects indirectly, although correspondence analysis scaled both objects and variables simultaneously. In this chapter, the primary focus is *Q*-mode analyses that directly scale objects based on similarities or dissimilarities between them. The techniques based on dissimilarities attempt to display the dissimilarities between objects graphically, with the distance between objects on the plot (inter-object distances) representing their relative dissimilarity. The scores for objects on the axes of these scaling plots can be used as variables in subsequent analyses so the techniques in this chapter are also methods for variable reduction. Remember that objects represent sampling or experimental units, such as plots, organisms, aquaria, sites etc..

Some of the dissimilarity measures for dichotomous and continuous variables were outlined in Chapter 15 (and see Legendre & Legendre 1998 for a much more complete treatment) and all of those dissimilarities can be used with the analyses in this chapter. However, the choice of dissimilarities is a crucial one and different dissimilarities can result in very different patterns in, and interpretations of, the analyses we will describe. Additionally, the form of transformation and/or standardization of variables and/or objects, combined with the dissimilarity measure, can also be very influential.

18.1. Multidimensional scaling

Multidimensional scaling (MDS) refers to a broad class of procedures that scale objects based on a reduced set of new variables derived from the original variables (Cox & Cox 1994). As the name suggests, MDS is specifically designed to graphically represent relationships between objects in multidimensional space. The objects are represented on a plot with the new variables as axes and the relationship between the objects on the plot should represent their underlying dissimilarity. The methods we described in Chapter 17 achieve this scaling indirectly, although MDS is more commonly based on similarities or dissimilarities between objects and was termed "similarities MDS" by Jackson (1991).

The basic data structure we will use in this chapter is similar to that from Chapter 17, a data matrix of *i* equals 1 to *n* objects by *j* equals 1 to *p* variables. Any two objects will be identified as *h* and *i* (*sensu* Legendre & Legendre 1998). The dissimilarities between objects calculated from our data are termed *d*, so that the dissimilarity between any two objects is d_{hi}. We will call the distance between any two objects (inter-object distances) in the scaling (configuration) plot $\tilde{d}_{hi}$ and it is usually measured as simple Euclidean distance. Unfortunately, there is some inconsistency in the symbols used for dissimilarity and inter-object distance in the literature, with *d* commonly used for dissimilarity. This seems inappropriate as Greek letters are usually reserved for unknown parameters.

MDS can be based on any of the measures of dissimilarity described in Chapter 15 but is not restricted to these. For example, Guiller *et al*. (1998) calculated genetic dissimilarities (Nei's and Rogers' indices) between 30 North African populations of the snail *Helix aspersa*, based on 17 enzyme loci. They used MDS to examine the relationships between the populations.

We will illustrate MDS using some recent data sets from the biological literature.

Genetic structure of a rare plant

In Chapter 15, we described the work of McCue *et al*. (1996), who measured the genetic structure of a rare annual plant (*Clarkia springvillensis*) in California. They identified eight subpopulations and calculated Cavalli-Svorza genetic distances between subpopulations from isozyme analysis of tissue samples. We will use their genetic

distances as dissimilarities and examine the relationships between the subpopulations using MDS.

Habitat fragmentation and rodents

In Chapter 13, we introduced the study of Bolger *et al*. (1997) who surveyed the abundance of seven native and two exotic species of rodents in 25 urban habitat fragments and three mainland control sites in coastal southern California. Besides the variables representing the species, other variables recorded for each fragment and mainland site included area (ha), % shrub cover, age (yr), distance to nearest large source canyon and distance to nearest fragment of equal or greater size. We will first calculate dissimilarities in species composition between the 25 fragments and three mainland sites and use MDS to represent the relationship between these objects. We will then examine relationships with other fragment characteristics such as area etc. for the 25 fragments.

Geographic variation and forest bird assemblages

Mac Nally's (1995) study on forest birds was first used in Chapter 17. The data set consisted of the maximum abundance (from four seasons) for 102 species of birds for 37 sites in southeastern Australia. These sites were actually replicates of five different forest types, four each of Gippsland manna gum, montane forest, foothills woodland, box-ironbark and river redgum with the remaining 17 sites not able to be classified into one of the habitats. An obvious question is whether the five habitat types are different in the composition of their bird assemblages.

18.1.1. Classical scaling – Principal Coordinates Analysis (PCoA)

Principal coordinates analysis (PCoA) is closely related to PCA (Chapter 17) and is sometimes called classical scaling. We will only provide a brief introduction to PCoA here (see Legendre & Legendre 1998 for complete details), mainly because it is not used that much as a scaling (ordination) technique in biology. The steps in PCoA are:

- Create an n by n matrix of dissimilarities between objects (d_{hi}), based on any of the dissimilarity measures described in Chapter 15.
- Transform these dissimilarities to $-0.5d_{hi}^2$. This transformation maintains the original dissimilarities during subsequent calculations (Legendre & Legendre 1998).
- These transformed dissimilarities are double centered by subtracting the means for the relevant row and column and adding the overall mean from the dissimilarity matrix. This centering removes the first, and trivial, eigenvector in the next step. The relative positions of the objects in the final configuration won't be affected by the double centering.
- This symmetric n by n matrix of transformed dissimilarities is then subjected to a spectral decomposition to obtain the eigenvectors and their eigenvalues, in exactly the same way as we treated a matrix of associations (covariances or correlations) between variables in a *R*-mode PCA. Most of the information (as measured by the eigenvalues) in the dissimilarity matrix will be in the first few eigenvectors (Box 18-1).
- As with PCA (Chapter 17), the eigenvectors are scaled, usually by the square roots of the eigenvalues (Legendre & Legendre 1998).
- The coefficients of these eigenvectors are these are used to position the objects relative to each other on the scaling plot (Figure 18-1).

If the original data were centered by variable means and Euclidean distance was used to create the matrix of dissimilarities between objects, the relative positions of objects in the PCoA scaling will be similar to those for the scaling plot from a PCA based on a matrix of covariances between variables. If the original data were double transformed by row

and column totals so that chi-square distance was used to create the dissimilarity matrix, the relative positions of objects in the PCoA scaling will be similar to those for the scaling plot from a CA. So PCoA can be viewed as a generalization of PCA that allows a much wider range of dissimilarity measures to be used.

Another way of viewing PCoA is a translation of dissimilarities between objects into Euclidean distances, actual distances between objects in multidimensional space (Legendre & Anderson 1999). If the original dissimilarities were metric (such as Euclidean or chi-square), and all eigenvectors are retained, then the distances in principal coordinate space are the same as the original dissimilarities because all the variance in the original dissimilarity matrix is retained in the principal coordinates. In contrast, biologists often use nonmetric dissimilarities, like Bray-Curtis for species abundance data, and the principal coordinates represent only part of the variation in the original dissimilarities. Unfortunately, the remainder may be represented by negative eigenvalues, which are very difficult to interpret. This may not be a problem if we are using PCoA as a variable reduction technique because the first few eigenvalues will be positive. However, if we wish to use all the principal coordinates derived from a nonmmetric dissimilarity matrix, such as in distance-based redundancy analysis (db-RDA; see Section 18.1.3), then we usually have to correct for the negative eigenvalues. These corrections are somewhat technical (Legendre & Legendre 1998, Legendre & Anderson 1999a) and may result in conservative tests of complex hypotheses (McArdle & Anderson (2001).

When dealing with species abundance data, Minchin (1987) showed that the scaling plots of sampling units produced by PCoA could distort underlying ecological gradients. In particular, PCoA would force long gradients (i.e. with considerable species turnover from one end to the other) into curved patterns in the configuration in second and higher dimensions. This distortion occurred even when more robust dissimilarity measures like Bray-Curtis were used and Minchin (1987) argued that this was because PCoA, like PCA, is based on a linear relationship between dissimilarity and ecological distance, whereas the relationship was nonlinear, particularly for large dissimilarities. Also, PCoA does not provide a simple way of interpreting the new coordinates in terms of the original variables (Legendre & Legendre 1998). While these problems do not rule out PCoA as a scaling technique for other types of data, biologists don't use PCoA very much by itself because modern desktop computers make enhanced scaling techniques (Section 18.1.2) so accessible. However, PCoA was used by Rundle & Jackson (1996) who measured the abundance of 15 species of littoral zone fish from five sites in each of three lakes in Ontario, Canada. They calculated Bray-Curtis dissimilarities between the 15 sites and then subjected the dissimilarity matrix to a PCoA. The first two axes explained over 69% of the variation in the original dissimilarity matrix and one lake clearly separated from the other two along the first axis.

We illustrate the use of PCoA on the data from Bolger et al. (1997), who recorded the abundance of nine species of rodents in 25 habitat fragments and three mainland sites in southern California – see Box 18-1. We calculated a matrix of Bray-Curtis dissimilarities between sites. Close to 90% of the variation was explained by the first two axes.

18.1.2. Enhanced multidimensional scaling

Enhanced algorithm

Methods for MDS more familiar to biologists involve additional steps, beyond the initial scaling used by PCoA, to improve the fit between the observed dissimilarities between objects (d_{hi}) and the inter-object distances in the configuration ($\tilde{d}_{hi}$). Jackson (1991) termed these methods "enhanced multidimensional scaling". Basically, these methods iteratively reposition the objects in the configuration using an algorithm that improves the fit between the dissimilarities and the inter-object distances, the latter measured by a form of Minkowski metric such as Euclidean distance. The most commonly used algorithm for

enhanced MDS is KYST, developed from methods first proposed by Kruskal (1964a,b), although some software offers the alternative ALSCAL program. The approach is surprisingly simple, although the computations would be very tedious without computer software. The steps for an enhanced MDS are (Figure 18-2):

1. Set up a data matrix and make decisions about transformations or standardizations of the data.
2. Calculate a matrix of dissimilarities between objects (d_{hi}) using any of the dissimilarities described in Chapter 15. Similarities could also be used; it makes no difference in the subsequent steps.
3. Decide on the number (k) of dimensions (i.e. axes) for the scaling, which will be a compromise between the need to get the fit between dissimilarities and inter-object distances as good as possible and minimizing the number of scaling dimensions for simple interpretation.
4. Arrange the objects in a starting configuration in the k-dimensional space (i.e. on the plot), either at random or more commonly using co-ordinates from a PCoA or even a PCA.
5. Move the location of objects in the k-dimensional space iteratively so that at each step, the match between the inter-object distances in the configuration ($\tilde{d}_{hi}$) and the actual dissimilarities (d_{hi}) improves. The iterative procedure uses the method of steepest descent (see Kruskal 1964a,b for details).
6. The final position of the objects and therefore the final configuration plot is achieved when further iterative moving of the objects can no longer improve the match between the inter-object distances in the configuration and the actual dissimilarities.

We can show the relationship between inter-object distance and dissimilarity for all pairs of objects in a Shepard diagram, which is simply a scatterplot with dissimilarity (d_{hi}) on the horizontal axis and inter-object distance ($\tilde{d}_{hi}$) on the vertical axis (Figure 18-2c). Now consider a linear or nonlinear regression model relating inter-object Euclidean distance ($\tilde{d}_{hi}$) as the response variable to dissimilarity (d_{hi}) as the predictor variable. The differences between the observed inter-object distances and those predicted by the regression model ($\hat{\tilde{d}}_{hi}$, sometimes termed "disparities" in the MDS literature) are the residuals from the regression model. These residuals can be used to measure the match between the calculated dissimilarities and the inter-object distances in the configuration. One measure of fit is Kruskal's stress:

$$\sqrt{\frac{\sum(\tilde{d}_{hi}-\hat{\tilde{d}}_{hi})^2}{\sum \tilde{d}_{hi}^2}} \qquad (18.1)$$

In equation 18.1, the summation is over all possible $n(n\text{-}1)/2$ pairwise distances and dissimilarities. If there is a perfect metric match between inter-object distance and dissimilarity (i.e. they are directly proportional to each other), then the residuals and stress will be zero. The lower the stress value, the better the match. There are other versions of stress used to measure fit (e.g. see Jackson 1991) and it is important to know which your software uses because they are scaled, and therefore interpreted, differently. The version in equation 18.1 is the one usually incorporated in the KYST algorithm and most commonly used by biologists. When stress is based on a parametric linear or nonlinear regression model relating inter-object distances to dissimilarities, we have metric MDS.

It is common for the Shepard plot to show a nonlinear relationship between inter-object distance and dissimilarity (Figure 18-3b). While this might suggest that a nonlinear

model relating inter-object distance and dissimilarity is most appropriate, a more robust approach is to fit a monotonic regression. This is a form of nonparametric regression that relates the rank orders of the two variables (Chapter 5). So stress now measures the concordance in the rank order of the observed inter-object distances and those predicted from the dissimilarities. When stress is based on rank orders, we have nonmetric MDS (NMDS).

A third type of MDS has been developed by Faith *et al*. (1987) and is termed hybrid MDS (HMDS). They noted that for species abundance data, sampling units at the ends of long ecological gradients often have few or no species in common and this can result in the nonlinear relationship between dissimilarity and inter-object ("ecological") distance mentioned in the previous paragraph. Importantly, it seemed that a linear relationship between dissimilarity and inter-object distance was appropriate for small dissimilarities but inappropriate for larger dissimilarities. Their hybrid approach generates two dissimilarity matrices. The first deletes dissimilarities above a threshold value and then uses a metric (linear) MDS to measure stress. The second matrix uses all the dissimilarities and uses a nonmetric MDS to measure stress. The final configuration is the one that minimizes the combination of the two stress values. The choice of dissimilarity threshold is a difficult one, with Faith *et al*. (1987) originally proposing 0.8 (for Bray-Curtis or Kulczynski dissimilarities) but also suggesting that some continuous function could also be used. Our experience is that HMDS does not offer much advantage over NMDS, even for ecological data sets, and is only available in specialized software anyway.

Interpretation of final configuration

We illustrate the use of NMDS with the data set on genetic differences between subpopulations of a species of plant from McCue *et al*. (1996) in Box 18-2, the habitat fragmentation study of Bolger et al. (1997) in Box 18-3 and the forest bird community study from Mac Nally (1995) in Box 18-4. The final configuration is the scatterplot of objects in a scaling or ordination diagram (Figure 18-3, Figure 18-4, Figure 18-5). The interpretation of this plot depends on how good a representation it is of the actual dissimilarities, i.e. how low the stress value is. Clarke (1993) provided some guidelines for stress values based on ecological (species abundance) data. Stress values greater than 0.3 indicate the configuration is no better than arbitrary and we should not try and interpret configurations unless stress values are less than 0.2, and ideally less than 0.1. These thresholds are for Kruskal's stress formula in equation 18.1, while some software may use different versions that require different guidelines. We can always reduce the stress value, i.e. improve the fit between dissimilarities and inter-object distances, by increasing the number of dimensions in the scaling. However, the more dimensions we use, the more difficult the display and interpretation of the final configuration, so we are trying to achieve a compromise between minimizing stress and minimizing the number of dimensions. Our experience with ecological data is that two or three dimensions will usually produce adequate configurations.

The final orientation of the configuration is arbitrary and it is only the relative distances between objects that are relevant to interpretation in MDS. It is preferable to rotate the final configuration so that the first axis lies along the direction of maximum variation. This can be achieved by a PCA on the MDS axis scores (Clarke & Warwick 1993) and will often be done automatically by MDS software. Note that actual values of the object scores are also arbitrary and these can be scaled in a number of ways; only the relative distances between the objects is important. Plots of the final configuration do not need scales on the axes as long as the axes are scaled identically.

Basically, the interpretation of final scaling (ordination) plot is subjective. Objects closer together are more similar (e.g. in species composition) than those further apart. A useful addition to the plot is a minimum spanning tree, where the objects are joined by lines so that the sum of line lengths is the smallest possible and there are no closed loops (Figure

18-5b). Minimum spanning trees can be applied to any scatterplot of points. For MDS configurations, objects joined by the shortest spans are closest on the plot and those separated by longest spans are furthest apart; the latter may separate different groups of objects (see Digby & Kempton 1987). Minimum spanning trees can be plotted in three dimensions, although they become ugly to interpret.

We may also have formal hypotheses we wish to test. For example, are dissimilarities between objects related to other differences, such as geographic distances? If the data consist of replicate objects within pre-defined natural (e.g. polluted area vs non-polluted area) or experimental (e.g. different nutrient treatments) groups, then we would probably test whether objects within a group are closer together than objects from different groups. Testing these hypotheses will be considered in Sections 18.1.3 and 18.1.5.

Convergence problems

The algorithms for enhanced MDS converge to the final configuration iteratively and the number of iterations depends on the complexity of the data. More rapid convergence can be achieved if the coordinates from an initial PCA or PCoA scaling are used rather than a random starting configuration and some software for MDS defaults to a preliminary PCoA before iterating. The iterative nature of the various algorithms for enhanced MDS means that the iterations can converge to a "local" solution that is not the configuration that best matches inter-object distances with dissimilarities. The only solution to this problem is to repeat the MDS a number of times, using a new random starting configuration each time, and then compare the different configurations for stress and axis co-ordinates. We can only be confident of the final configuration if it occurs from a majority of random starts. Comparison of different configurations can be achieved through Procrustes analysis (Digby & Kempton 1987), where one configuration is rotated and rescaled to most closely match a second configuration of the same objects. The fit is measured by the sum of squared distances between the corresponding objects in the two configurations.

18.1.3. Dissimilarities and testing hypotheses about groups of objects

It is common for biologists to have recorded multiple variables from objects in a sampling or experimental design where the objects fall into pre-defined groups. The design might have a single factor or be multifactorial with factors either crossed or nested. We would often be interested in testing null hypotheses about differences between groups in these designs, as we would using linear ANOVA models if we had just a single response variable. In the multivariate context, the methods for testing such hypotheses proposed in the literature are based on the original variables, the scores for each object in scaling (ordination) space or the dissimilarities between objects.

Tests based on dissimilarities are not straightforward for two reasons. First, the dissimilarities between objects are not always independent of each other (the dissimilarities between objects 1 and 2 and 2 and 3 are not independent of the dissimilarity between objects 1 and 3), so randomization (permutation) testing procedures are required (Chapter 3). Second, if we wish to use the dissimilarities in linear models, we require sums-of-squares based on the difference between each observation and the mean of the observations, or the centroid in the multivariate context. When dealing with metric dissimilarities (e.g. Euclidean distance), the centroid of a group of observations and the sum of squared deviations from this centroid are straightforward to calculate and interpret. This is not the case when dealing with nonmetric dissimilarities like Bray-Curtis and a limitation of some approaches is their inability to deal with nonmetric dissimilarities (see Anderson 2001).

MANOVA based on original variables

We could use a multivariate analysis of variance (MANOVA; see Chapter 16), a multivariate analogue of the univariate ANOVA, to test the null hypothesis of no

difference between groups in some linear combination of variables. While MANOVA may be useful in some situations, it has quite restrictive assumptions about variances and covariances that are difficult to test (Chapter 16) and are unlikely to be met when the variables are species abundances with lots of zeros. A robust nonparametric form of MANOVA (NPMANOVA) that uses dissimilarities has recently been described by Anderson (2001) and will be discussed below. MANOVA comparing groups of objects is also restricted to data sets where the number of variables does not greatly exceed the number of objects, whereas ecological data sets often comprise many variables (species) and fewer objects (sampling untis).

(M)ANOVA based on axis scores

Another approach is to use any of the scaling procedures from Chapter 17 or this chapter that provide scores for each object on derived variables (components or axes). These scores could be used as response variables in linear models, as described for PCA in Chapter 17, to test hypotheses about group differences. There are some problems with this method. With MDS, we have to decide which axes to use; maybe scores from multiple axes (i.e. the first 2 or 3 dimensions if stress is adequate) could be used with a MANOVA? The axes themselves are also not a linear combination of variables like the components from a PCA or axes from a CA so are more difficult to relate to the original variables. Finally, the MDS axes simply define the relative positions of the objects in multidimensional space so as to represent the observed dissimilarities. Tests of hypotheses about group differences might be better based on these actual dissimilarities rather than some approximation of them.

Mantel test

The Mantel test described in Chapter 15 can be used to correlate a dissimilarity matrix between objects with another dissimilarity matrix that simply separates objects into groups (Manly 1997, Schnell *et al*. 1985). This second matrix is termed the model or design matrix (Legendre & Legendre 1998, Sokal & Rohlf 1995). The main limitation of using the Mantel test in this way is that it is difficult to test more complex models such as those including interaction terms.

Rundle & Jackson (1996) used a Mantel test to test for differences in the fish communities of the littoral zones of three lakes in Canada based on five sites in each lake. They constructed a Bray-Curtis dissimilarity matrix between the 15 sites. To test whether the variation in fish communities was primarily between lakes rather than within lakes, they used Mantel test to assess whether the Bray-Curtis matrix based on fish was associated with a matrix containing zeros for within lake distances between sites and ones for between lake distances between sites.

Multi-response permutation procedures

Mielke *et al*. (1976) proposed multi-response permutation procedures (MRPP) that test hypotheses about group differences in Euclidean distances and Zimmerman *et al*. (1985) illustrated their application to biological data sets, such as *n* sampling units by *p* species. Basically, the MRPP determines the mean of the Euclidean distances between objects within each group and calculates an MRPP statistic (delta) that is a linear combination of these mean within-group Euclidean distances. The statistic produces a weighted average (based on sample size) of the within-group mean Euclidean distances. Small values of the statistic indicate that objects tend to be found in groups. The probability distribution of the MRPP statistic is determined by randomizing the allocation of all objects to the groups, keeping the original sample sizes, with the null hypothesis being that all random allocations are equally likely. We compare our observed value of the MRPP statistic to the probability distribution generated under randomization to get the probability of obtaining the observed value of the statistic or one smaller under the null hypothesis. The

MRPP can be used for a range of hypotheses including those associated with paired comparisons and randomized block designs.

MRPPs have been traditionally based on Euclidean distance and their use with more robust nonmetric dissimilarities would be tricky because of the difficulty of defining the centroid and calculating the mean within-group dissimilarity. Nonetheless, McCune & Mefford (1999) have suggested that MRPPs might work well with other dissimilarity measures, such as Bray-Curtis. Since Euclidan distance is not a particularly appropriate measure of dissimilarity for some types of biological data, e.g. species abundances (Chapter 15), we could use the inter-object distances from classical (PCoA) or enhanced scaling (NMDS) in a MRPP. This is not an ideal solution because we know that these distances are an imperfect representation of the actual dissimilarities, and correction for negative eigenvalues would be required for PCoA. This approach is used, although not for MRPP, in distance-based redundancy analysis (Anderson & Legendre 1999a) and discussed below.

Analysis of similarities

ANOSIM (Analysis of Similarities; Clarke 1993, Clarke & Warwick 1994) is a hypothesis testing procedure that uses Bray-Curtis dissimilarities, although it could use any dissimilarity measure. This procedure uses a test statistic (*R*) based on the difference between the average of all the rank dissimilarities between objects between groups ($\bar{r}_B$) and the average of all the rank dissimilarities between objects within groups ($\bar{r}_W$):

$$R = \frac{\bar{r}_B - \bar{r}_W}{n(n-1)/4} \qquad (18.2)$$

This is analogous to an ANOVA comparing between- and within-group variation. The use of rank dissimilarities rather than actual dissimilarities is in keeping with the spirit of nonmetric MDS.

The H_0 being tested by ANOSIM is that the average of the rank dissimilarities between all possible pairs of objects in different groups is the same as the average of the rank dissimilarities between pairs of objects in the same groups. *R* is scaled to be within the range +1 to –1. Differences between groups would be suggested by *R* values greater than zero where objects are more dissimilar between groups than within groups. *R* values of zero indicate that the null hypothesis is true. Negative *R* values indicate that dissimilarities within groups are greater than dissimilarities between groups, an outcome Clarke & Warwick (1994) considered unlikely. However, Chapman & Underwood (1999) showed that negative *R* values can occur, especially when groups had high levels of within-group variability that were similar between groups and when outliers were present. They argued that negative *R* values could be a useful diagnostic, indicating an inappropriate completely random sampling design when stratified sampling would be appropriate.

Like the MRPP, ANOSIM uses a randomization procedure to randomly allocate objects to groups to generate the distribution of *R* under the null hypothesis that all random allocation are equally likely. Clarke & Warwick (1994) described the use of ANOSIM procedures for nested designs where averaging over the subsampling levels produces a series of single factor tests for each factor. They also proposed ANOSIM for testing main effects in factorial designs by simply treating each main effect as a single factor test, averaging over the other factor. Legendre & Legendre (1998) pointed out that ANOSIM is very similar to a Mantel test using a model matrix to define the groups specified in the hypothesis and the two methods should produce similar *P* values for the same hypothesis.

Both MRPP and ANOSIM use some measure of average dissimilarity within and between groups. Van Sickle (1997) described a useful graphical display for representing the relative strength of the differences in dissimilarity between groups, called a mean similarity dendogram. In its simplest from, a mean similarity dendogram for two or more

groups would have branches for each group originating at the between-group mean dissimilarity and the length of each branch representing the within group mean dissimilarities. Alternatively, the origin of each group branch could be staggered, with the mean between group dissimilarity for each pair of groups plotted separately. Displays for multifactor designs are also possible (Van Sickle 1997). Mean similarity dendograms use the actual mean dissimilarities, rather than their rank orders, for plotting and therefore do not provide a direct graphical representation of the ANOSIM results.

One of the limitations of both MRPP and the ANOSIM procedure is that complex tests, such as interaction terms in linear models, are not available. This is in part because tests of interactions are difficult in the randomization context, since the interaction hypothesis cannot be simply expressed in terms of a random reallocation of observations to groups (see slightly differing opinions in Edgington 1995 and Manly 1997). Interactions are most sensibly tested in a linear model framework that also considers main effects. Unfortunately, if nonmetric dissimilarities like Bray-Curtis are used, it is not straightforward to partition the variance (sum-of-squares) from fitting a multivariate linear model because of the difficulty of defining deviations from the centroid of the observations (Anderson 2001, Legendre & Anderson 1999a).

Distance-based redundancy analysis

Because of the difficulties in using MRPP or ANOSIM tests for designs with interactions, Legendre & Anderson (1999a; see 1999b for minor correction) proposed an alternative approach for testing group differences in dissimilarities, called distance based redundancy analysis (db-RDA). Their method uses PCoA to convert the original dissimilarities into their equivalent Euclidean distances, correcting for negative eigenvalues (Section 18.1.1). The matrix of n objects by p principal coordinates is then related to grouping factors using redundancy analysis (RDA; Chapter 17), where the grouping factors are represented by a matrix of dummy variables (Chapter 5) and the relationship is tested by a linear model using randomization tests (Chapters 3 and 8). This makes it easy for testing interactions because the analysis just becomes a multiple linear regression model and any combination of crossed and nested, fixed and random factors can be included.

It turns out that we can get the same results by simply doing a MANOVA test on the corrected principal coordinates, although Legendre & Anderson (1999a) argued that db-RDA has the advantages of more robust randomization tests and does not require more objects than variables in the original data matrix. The latter advantage is important because ecological data sets nearly always have more species (variables) than sampling units (objects). The main limitation of db-RDA is its complexity and the need to have software for the RDA component.

Nonparametric MANOVA

Distance-based RDA was developed to translate various nonmetric measures of dissimilarity into their equivalent distance in Euclidean space using PCoA. We can then relate these distances to a design matrix using linear models (e.g. RDA) and calculate sum-of-squared deviations between observations and their centroid. McArdle & Anderson (2001) and Anderson (2001) have recently shown that the partitioning of sums-of-squares (SS) and variances used for testing linear models can also be applied directly to dissimilarities, even nonmetric ones like Bray-Curtis. This method means that using PCoA on the original dissimilarities is not necessary and the negative eignevalues produced by db-RDA correspond to negative SS. The correction for negative eigenvalues in db-RDA described by Legendre & Anderson (1999) actually produces overly conservative tests when random factors are included in the design (McArdle & Anderson 2001).

The nonparametric MANOVA described by McArdle & Anderson (2001) and Anderson (2001) is elegantly simple and can be applied to any design structure. The main difficulty is developing a randomization test for complex terms like interactions (Chapter 9; see

Manly 1997). Our view is that the nonparametric MANOVA is so widely applicable in the biological sciences that we will describe it in some detail.

Consider a single factor design with p groups and n objects in each group so the total number of objects is $N = pn$. For the equations below, any two objects are termed h (h = 1 to N) and i (i = 1 to N). From an N by N matrix of dissimilarities (d_{hi} e.g. Bray-Curtis) between all pairs of objects, we calculate three SS.

The first is the sum of squared dissimilarities between all pairs of objects divided by N:

$$SS_{Total} = \frac{1}{N}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N} d_{hi}^2 \qquad (18.3)$$

Note that only the lower (or upper) diagonal of the dissimilarity matrix is used. The dissimilarity between objects h and i is the same as between i and h and is only counted once in the calculation of SS_{Total}.

The second is the within-groups SS. The $SS_{Residual}$ is the sum of squared dissimilarities between objects with each group, summed over the groups:

$$SS_{Residual} = \frac{1}{n}\sum_{h=1}^{N-1}\sum_{i=h+1}^{N} d_{hi}^2 e_{hi} \qquad (18.4)$$

In equation 18.4, e_{hi} equals one if object h and i are in the same group and zero if they are in different groups (just the design matrix in the Mantel test above).

The between-groups SS is determined from the usual additive partitioning of the total SS described for ANOVA models in Chapter 8:

$$SS_{Groups} = SS_{Total} - SS_{Residual} \qquad (18.5)$$

The approximate F-ratio statistic for testing the H_0 that all allocations of objects, and therefore dissimilarities between objects, between groups are equally likely is:

$$F = \frac{SS_{Groups}/(p-1)}{SS_{Residual}/(N-p)} \qquad (18.6)$$

This is analogous to the F-ratio statistic for a single factor ANOVA model. The randomization test is then done in the same manner as described for single factor ANOVA tests in Chapter 8, using a subset of all possible permutations for anything except very small p and n.

Pairwise contrasts of specific groups, either planned or unplanned, can be done using the same test statistic. If there are many contrasts, the significance levels may need to be adjusted to control familywise Type I error rate, using one of the Bonferroni corrections described in Chapter 3.

However, the main advantage of this nonparametric MANOVA is that it can handle more complex designs, especially those that include interactions. Anderson (2001) provides appropriate formulae for factorial designs but the logic is straightforward. The SS_{Total} are calculated using equation 18.3. The main change from a single factor design is that we need to calculate within-group SS for each factor separately, ignoring the other factor. The SS for each main effect are simply the difference between the SS_{Total} and within-group SS for that factor. The $SS_{Residual}$ are calculated using equation 18.4 except that each combination of the two factors (each cell) is considered a single group. So the e_{hi} equal one if the objects are in the same cell (combination of factors) and zero if they are in different cells. The $SS_{Interaction}$ are what is left after the main effects and residual SS are subtracted from the total. The F-ratios are determined following equation 18.6, although the denominator may need to be changed if either factor is random (see Chapter 9).

As we discussed in Chapter 9, there are different approaches to randomization tests in factorial designs and some debate about whether randomization tests for interaction terms

are possible. Manly (1997) summarized these different approaches, including whether to randomize observations or residuals and whether to impose restrictions on which objects are randomized for tests of different terms. He argued that the different methods produced comparable results.

We illustrate the use of a single factor nonparametric MANOVA with the bird community data from Mac Nally (1995) – see Box 18-4. There were four replicate sites for each of five forest habitats types; unclassified sites were not included in the comparison. There was a significant difference between habitats, although like the ANOSIM procedure, the small number of possible permutations with only four replicates per group meant that pairwise comparisons were difficult to interpret after adjusting significance levels. Based on raw *P* values, the nonparametric MANOVA procedure seemed more powerful than the ANOSIM comparisons.

The two main advantages of the nonparametric MANOVA introduced by McArdle & Anderson (2001) and Anderson (2001) are that any dissimilarity measure can be used and the tests are based on the partitioning of sums-of-squares as used in classical linear models. This means that the method can be used for any design structure that can be formulated as a linear model (see Chapters 5, 6, 8 to 12) and can accommodate fixed and random factors by using different denominators in the approximate *F*-ratios. The only limitation is the difficulty of determining the appropriate randomization test procedure for complex designs.

18.1.2. Relating MDS to original variables

Another question of interest in scaling (ordination) procedures is to determine which variables contribute most to the observed pattern among objects, e.g. which species contribute most to the separation among sampling units or which morphological variables contribute most to the separation of organisms. As described in Section 18.1.3, we will often be using a sampling or experimental design that includes groups of objects and our interest will be which variables contribute most to the any separation among groups. When we scale using one of the *R*-mode methods described in the previous chapter, then we obtain loadings for each variable on each derived component (axis of the scaling plot) as in PCA or can plot object and variable scores jointly to examine correlations as in CA.

Scaling techniques that are based directly on dissimilarities, such as MDS, do not provide correlations between derived axis scores and variables as part of the algorithm but there are alternative ways of investigating how the variables contribute to the final configuration of objects. We could simply correlate the axis scores from an MDS with each variable or linear combination of variables. This is not an ideal solution because, besides the problem of increasing Type I error rates from multiple testing if we do numerous correlations, we have to decide how many and which dimensions from the MDS we use. Additionally, we know that the scores, or at least the distances between objects, are imperfect representations of the actual dissimilarities so a method that uses these dissimilarities directly would be preferable.

Clarke & Warwick (1994) described a procedure for ecological data termed SIMPER (similarity percentages) for determining which species (variables) are contributing most to the dissimilarity between groups of object (sampling units). For example, the Bray-Curtis dissimilarity for a pair of sampling units is basically the differences between the units for each species, summed over all the species. SIMPER computes the % contribution of each species to the dissimilarities between all pairs of sampling units in different groups and the % contribution of each species to the similarities between all pairs of sampling within each group. It then calculates the average of these % contributions with standard deviation. Species with a large ratio of average / standard deviation % contribution to dissimilarity between sampling units in different groups are those species that best discriminate between the groups. Note that there are no formal tests of hypotheses with SIMPER, just a list of species in order of their % contributions to dissimilarities between groups or similarities within groups.

18.1.3. Relating MDS to covariates

In ecological data sets, we often have two types of variable recorded for each sampling unit, species abundances (presence/absence) and environmental characteristics (covariates). In these circumstances, we might wish to relate the dissimilarities between sampling units, or groups of sampling units, based on the species variables to differences in the environmental characteristics. Are sampling units that are very different from others in terms of species composition also very different in terms of one or more environmental variables? There are numerous ways of relating dissimilarities between sampling units to environmental variables, two of which we have already described. We could examine correlations between, or fit regression models to, the scores for each axis from the MDS and the environmental variable(s) (Ludwig & Reynolds 1988), just as we described for component scores from a PCA in Chapter 17. These correlations can be represented as vectors on the MDS plot, producing a biplot, and tests of the correlations are best done in a randomization context. The problems with relating environmental variables (covariates) to axis scores are the same as outlined in Sections 18.1.3 and 18.1.4, i.e. the problem of multiple testing, axis scores being an imperfect representation of the actual dissimilarities, deciding how many and which dimensions to use.

Clarke & Ainsworth (1993) proposed a procedure for ecological data that basically measures the correlation between dissimilarities between sampling units based on species composition and the dissimilarities between sampling units based on environmental variables. They provided an algorithm called BIO-ENV that first calculates a dissimilarity matrix (e.g. Bray-Curtis) between sampling units based on species abundances and a separate dissimilarity matrix (e.g. Euclidean distance) between sampling units based on environmental variables. It then measures any correlation between the rank-orders of these two matrices using the Spearman rank correlation coefficient. Each pair of observations for the correlation will be the rank of the Bray-Curtis dissimilarity (from species abundances) between objects *h* and *i* and the rank of the Euclidean distance (from environmental variables) between objects *h* and *i*.

Legendre & Legendre (1998) pointed out that the BIO-ENV procedure basically calculates the same correlation as a Mantel test (Chapter 15 and Section 18.1.3), except the former is based on rank transformed data. The Mantel test could be used for the global test of no correlation between the two matrices, or even between the dissimilarities based on species composition and differences between sampling units for each environmental variable separately. It can also be extended to compare more than two matrices (Diniz-Filho & Bini 1996).

Clarke & Ainsworth (1993) and Clarke & Warwick (1994) incorporated a stepwise routine into their BIO-ENV procedure, to find the combinations of environmental variables that produce dissimilarities between sampling units with the highest correlations with dissimilarities between sampling units based on species composition. They argued that their implementation of the Mantel test is not suitable for hypothesis testing, both because the dissmilarities for both sets of variables are not independent and also because their stepwise procedure would produce numerous significance tests that are difficult to interpret (see Chapter 6).

Procrustes analysis (Section 18.1.2; Digby & Kempton 1987, Legendre & Legendre 1998) can also provide a descriptive measure of the fit of a configuration between objects based on one set of variables (e.g. species abundances) and a configuration between the same objects based on a separate set of variables (e.g. environmental characteristics).

18.2. Classification

The aim of classification is to group together a number of objects based on their attributes or variables to produce groups of objects where each object within a group is more similar to other objects in that group than to objects in other groups. One form of classification analysis is discriminant function analysis (DFA; Chapter 16) where the

number of groups was known *a priori*. In this section, we are interested in classification methods where the number of groups is not known and must be determined from the data.

18.2.1. Cluster Analysis

Cluster analysis is a method for combining similar objects into groups or clusters, which can usually be displayed in a tree-like diagram, called a dendogram (Figure 18-6a). Legendre & Legendre (1998) provide a recent, very thorough, discussion. Cluster analyses are used commonly by biologists. For example, Crews *et al*. (1995) examined plant species in montane rainforest in Hawaii. They compared six sites (varying in age) using the cover-abundance measures for numerous plant species. The objects were sites, the variables were species abundances and cluster analysis was used to place the sites into like groups. Koenig *et al*. (1994) studied acorn production in oak trees in California. They clustered five species of oaks (objects) based on twelve mean annual values of acorn production (variables). Probably the most important use of cluster analysis in biology is taxonomic and phylogenetic research, where the dissimilarity measures are often morphological or genetic/molecular differences between organisms, species etc. and the dendogram represents a possible evolutionary sequence.

Agglomerative hierarchical clustering

Agglomerative methods start with individual objects and join objects and then objects and groups together until all the objects are in one big group. This is the form of cluster analysis familiar to most biologists. Usually objects are clustered but sometimes you may wish to cluster variables (e.g. species). Most algorithms for agglomerative cluster analysis start with a matrix of pairwise similarities or dissimilarities between the objects and the steps are as follows:

1. Calculate a matrix of dissimilarities (d_{hi}) between all pairs of objects.
2. The first cluster is formed between the two objects with the smallest dissimilarity.
3. The dissimilarities between this cluster and the remaining objects are then recalculated.
4. A second cluster is formed between cluster 1 and the object most similar to cluster 1.
5. The procedure continues until all objects are linked in clusters.

The graphical representation of the cluster analysis is a dendogram (Figure 18-6a, Figure 18-7), showing the links between groups of objects with the lengths of the lines representing dissimilarity. If there are many objects, the standard dendogram can be very long and difficult to represent on a single page. An alternative representation is the polar dendogram (Figure 18-6b), where the objects are arranged in a circle and their distance from the centre of the circle represents dissimilarities between objects and groups of objects. Like scaling (ordination) plots, the interpretation of the groupings in the dendogram is subjective and the decision about which groups to report is usually based on some arbitrary cut-off value for dissimilarity.

The major difference between the variety of available hierarchical agglomerative clustering methods is how the dissimilarities between clusters and between clusters and objects (step 3) are recalculated. These are termed linkage methods and three common ones are:

- Single linkage (nearest neighbour), where the dissimilarity between two clusters is measured by the minimum dissimilarity between all combinations of two objects, one from each cluster.
- Complete linkage (furthest neighbour), where the dissimilarity between two clusters is measured by the maximum dissimilarity between all combinations of two objects, one from each cluster.

- Average linkage (group average or mean), where the dissimilarity between two clusters is measured by the average of all the dissimilarities between all combinations of two objects, one from each cluster. The group mean (or average) linkage strategy commonly called unweighted pair-groups method using arithmetic averages (UPGMA) is often recommended. There is a weighted version of UPGMA (WPGMA), which weights the original dissimilarities differently, and unweighted clustering based on centroids (UPGMC), which is equivalent to UPGMA except that centroids instead of means are used.

Kent & Coker (1992), Legendre & Legendre (1998) and Ludwig & Reynolds (1988) discuss the pros and cons of these different linkage methods. If there are "strong" (i.e. very dissimilar) groups in your data, then the different methods will produce similar dendograms; in contrast, the different linkage strategies can produce very different patterns for data with weak structure (Ludwig & Reynolds 1988). Belbin *et al*. (1993) proposed a flexible modification of UPGMA that allowed the clusters to be better, if artificially, defined and this method effectively recovered true groups in the data based on simulation studies (Belbin & McDonald 1993).

Box 18-5 illustrates a cluster analysis of the subpopulations of *Clarkia springvillensis* based on genetic differernces recorded by McCue et al. (1996). A cluster analysis of the 37 sites in southeastern Australia, using Bray-Curtis dissimilarities based on the densities of 102 species of forest birds (Mac Nally 1995), is presented in Box 18-6.

Agglomerative cluster analysis does have some disadvantages, primarily related to the interpretation of the dendogram. The hierarchical approach means that once a group or cluster is formed from two or more objects, that group cannot be broken later in the process. As a result, the dendogram is not a representation of all pairwise dissimilarities between objects like in multidimensional scaling (MDS). A misleading cluster formed early in the process will influence the remaining clusters. Also, the analysis forces objects into clusters and it would be easy for naïve biologists to place too much emphasis on these clusters without examining the actual dissimilarities. We much prefer MDS as a method for graphically representing relationships between objects based on dissimilarities.

Divisive hierarchical clustering

Divisive methods have a long history for clustering ecological data. They basically start with the objects in a single group and split them up into smaller and smaller groups until each group is a single object. One method popular with ecologists is two way indicator species analysis (TWINSPAN), a complex procedure that uses the reciprocal averaging algorithm of correspondence analysis (Chapter 17) to successively divide the first axis for both sampling units and species into smaller groups. The output includes a two-way table that orders the sampling units and species and shows the groupings and the relative abundances of species for each sampling unit. The actual computations are tedious, although a detailed description can be found in Kent & Coker (1992). Van Groenewood (1992) and Belbin & McDonald (1993) provided simulation results that showed that TWINSPAN is not particularly good at detecting true clusters in ecological data and the problems that affect correspondence analysis, particularly the distortion of sampling units along the first axis, also affect TWINSPAN.

Non-hierarchical clustering

Non-hierarchical methods do not represent the relationship between objects in hierarchical form. Basically, they start with a single object and cluster other objects that are similar to the first one. In contrast to hierarchical clustering, objects can be reassigned to clusters during the clustering process. One method common in statistical software is *K*-means clustering – see Legendre & Legendre (1998) for a detailed description. *K*-means works by splitting the objects into a pre-defined number (*K*) of clusters, and then cluster membership of objects is iteratively re-evaluated by some

criterion, such as to maximize the ratio of between-cluster to within-cluster variance. Another method is additive tree clustering, which develops a tree-like network (dendogram) where the dissimilarity between objects within a cluster is represented by the sum of the lengths of the branches joining them (Gower 1996) and may be more suited to nommetric dissimilarity measures.

18.3. Scaling (ordination) and clustering for biological data

When the main purpose of the multivariate analysis is to scale objects, what ecologists term ordination, numerous techniques are available. There have been many evaluations and comparisons of these techniques, particularly for ecological data in the form of species abundances across sampling units. Differing opinions on the relative merits of different techniques can be found in Faith *et al*. (1987), Jackson & Somers (1991), Minchin (1987), Palmer (1993), Peet *et al*. (1988), ter Braak & Verdonschot (1995), van Groenewood (1992), and Wartenberg *et al*. (1988), among others. In our view, the choice of method depends on the nature of the data, the implicit measure of dissimilarity used by each method, and not surprisingly, the biological question being addressed. Our preferred approach is to use a method that is applicable to a range of data types, is amenable to various user-defined standardizations and transformations of the data, is flexible in terms of which dissimilarity measure is used, and can be used for describing patterns and testing *a priori* hypotheses. Multidimensional scaling (MDS), especially the robust nonmetric version (NMDS), meets all these criteria. Any measure of dissimilarity can be used, thereby allowing dissimilarities between objects based on continuous, binary and mixed variables under nearly every combination of transformation and standardization. The scaling or ordination has been shown to be robust for a range of data types, accurately representing underlying true dissimilarities and recovering ecological gradients, and hypothesis tests can be based on the dissimilarities. For ecological data, NMDS also appears to be the most robust for nonlinear relationships of species abundances across sampling units along long ecological gradients, which can result in misleading arching of second and higher dimensions in some methods.

The most obvious competing technique is correspondence analysis (CA) or the more sophisticated canonical version (CCA). The strengths of these methods are also their weaknesses. By implicitly using the chi-square metric as the dissimilarity measure, they allow joint scaling plots of objects and variables and when axes are scaled similarly, relative positions of objects and variables can be compared. Unfortunately, the restriction to the chi-square metric also reduces flexibility and this dissimilarity measure may not be ideal for some forms of data (Faith *et al*. 1987). There are also decisions to be made about how to scale the axis scores, although the different scalings don't often alter the general pattern from the joint plot.

Constrained ordinations like CCA and redundancy analysis (RDA) also allow for biplots, where covariates can be included on the scaling plot showing which axes are correlated with which covariates. This is probably the main reason for the popularity of these methods, especially CCA. Relationships between dissimilarities and covariates under the MDS framework can also be evaluated although not in the same direct manner as in CCA and RDA. Finally, we shouldn't forget the oldest of these techniques, principal components analysis (PCA). While not always suitable as a scaling/ordination procedure, PCA is still a very important method for variable reduction, especially when linear relationships between variables are expected.

You may have inferred from Section 18.2.1 that we are not big users of cluster analysis, especially for representing dissimilarities between objects. Clustering procedures do not really use all pairwise dissimilarities for grouping objects so the dendogram is not necessarily a good representation of a dissimilarity matrix. The main use of clustering procedures in biology is to display possible evolutionary and phylogenetic relationships,

where the objects are organisms or taxonomic groups and the dissimilarities are morphological or genetic differences. Cluster analysis has less applicability for analyzing species abundance data to show relationships among sampling units. Ecologists sometimes use an initial cluster analysis to identify groups in a data set and then indicate those groups on a subsequent scaling plot. This approach has never made much sense to us, the cluster analysis almost certainly being a less efficient way of representing dissimilarities between objects than a method like enhanced MDS (but see Legendre & Legendre 1998 for an alternative view). Certainly, it is inappropriate to test hypotheses about differences between these groups; hypothesis tests cannot be validly used to compare groups that were defined by the same data.

18.4. General issues and hints for analysis

General issues

- Principal coordinates analysis (PCoA) is a useful metric scaling procedure but has generally been superseded by enhanced, iterative scaling procedures.
- Our preferred technique for scaling or ordination of ecological data, when there are numerous zeros and extracting underlying ecological gradients is important, is a combination of a suitable dissimilarity measure, like Bray-Curtis, and robust nonmetric multidimensional scaling.
- Nonmetric MDS is probably more robust than metric MDS, especially when the relationship between dissimilarities and inter-object distances is nonlinear. Hybrid MDS may offer a slight advantage.
- Hierarchical cluster analysis is not as useful as MDS for representing a dissimilarity matrix and has the disadvantage of forcing all objects into clusters that cannot be reassessed during the clustering procedure.

Hints for analysis

- Final enhanced MDS configurations should not be interpreted without examining stress values. Make sure you know which version of stress your software uses. Values for version one of Kruskal's stress should be less than 0.15, ideally less than 0.10, for configurations of objects to be considered reliable.
- Multiple runs from random starting configurations should be compared with enhanced MDS, to ensure that any configuration does not represent a local, unrepeatable, pattern. With large data sets, i.e. many objects, using an initial PCoA to determine a starting configuration may help convergence.
- Analysis of similarities (ANOSIM) or multi response permutation procedures (MRPP) are a useful ways of testing hypotheses about group differences in a multivariate context, the former retaining the underlying philosophy of NMDS. For pairwise comparisons of groups, *n* greater than four per group is needed for the randomization tests. For more complex hypotheses, especially tests of interactions, the nonparametric MANOVA of Anderson (2001) offers great promise.
- The unweighted pair-groups method using arithmetic averages (UPGMA) is usually recommended as a linkage strategy for agglomerative clustering. Non-hierarchical methods may offer more flexibility because clusters are not fixed once formed.

Box 18-1 Worked example of PCoA: habitat fragmentation and rodents.

We will use the data on rodent numbers from 25 canyon fragment and three mainland sites in California from Bolger *et al*. (1997) to illustrate PCoA. Because the sites were very different in size, we standardized the total abundance for each site to range between zero and one and calculated a matrix of Bray-Curtis dissimilarities between the sites. This matrix was then used for the PCoA. Of the 28 possible eigenvectors, ten had zero eigenvalues and seven had negative eigenvalues but nearly 90% of the variance was explained by the first two components so only these were used for the scaling plot of sites.

	Axis 1	Axis 2
Eigenvalues	5.255	1.724
% variation	66.081	21.681
Cum % variation	66.081	87.762

The PCoA scaling plot of the 28 sites based on the original Bray-Curtis dissimilarities of data range standardized by site is shown in Figure 18-1. When corrected for total abundance at a site, the three mainland sites were almost identical and were not distinguishable from most of the canyon fragments. Acuna, El Mac and 54th Street separated from the other sites, especially along axis 2. These three sites also stood out from the others in the scaling plot from a CA of these 28 sites (Chapter 17, Figure 17.5). The agreement with CA is because the latter emphasizes proportional abundance of species at each site, as does the PCoA when the dissimilarity is calculated on abundances standardized to the same maximum value at a site. Note, however, that the CA did separate the three mainland sites from each other, a pattern not observed in the PCoA, probably reflecting differences in the sensitivity of the two dissimilarity measures (chi-square and Bray-Curtis) to changes in proportional abundance.

Box 18-2 Worked example of enhanced MDS: genetic structure of a rare plant.

McCue *et al*. (1996) sampled eight subpopulations of the rare annual plant (*Clarkia springvillensis*) from three sites along the Tule River in California. Two sites, Bear Creek (BC) with three subpopulations and the Springville Clarkia Ecological Reserve (SCER) with three subpopulations, were separated by about 300m and the third site, Gauging Station (GS) with two subpopulations, is approximately 8 km apart. The nonmetric MDS algorithm produced identical configurations from all random starts and the stress of the final configuration was 0.045, indicating that the scaling/ordination of the subpopulations closely matched the Cavalli-Svorza genetic distances between the subpopulations. The final scaling plot of the subpopulations (Figure 18-3) indicates that the two Gauging Station subpopulations are genetically different from the remaining subpopulations, with subpopulation GS1 being the most distinct.

Box 18-3 Worked example of enhanced MDS: habitat fragmentation and rodents.

We will use the data on rodent numbers from 25 canyon fragment and three mainland sites in California from Bolger *et al*. (1997) to illustrate NMDS. Because the sites were very different in size, the data were standardized so that each site had a maximum total abundance of rodents of one. We were interested in comparing sites based on species composition and abundance but without patterns being confounded by very different areas.

A matrix of Bray-Curtis dissimilarities between all 28 sites was calculated and subjected to nonmetric MDS. From 20 random starts in two dimensions, the minimum stress value of 0.054 was achieved from 4 starts, although all 20 starts produced very similar final configurations, one of which is displayed in Figure 18-4, with a small range of stress values (0.054 to 0.059). The final configuration is shown in Figure 18-4. The mainland sites were not clearly separate from the fragments and the pattern of sites was similar to that in the PCoA plot. The same fragment sites were close to the mainland sites and Acuna, El Mac and 54th Street were most different to the mainland sites (Figure 18-4). It is interesting to compare the pattern from the NMDS to that from the CA on the same data described in Chapter 17 (Figure 17.5). Although the distances between the sites are different in the two plots, the broad pattern of Acuna, El Mac and 54th Street being separate was consistent in both analyses.

Correlations were calculated between the two dimensional configuration (scores) of sites and each of the six habitat variables (total area, shrub area, % area of shrubs, distance to nearest large source canyon and distance to nearest fragment of equal or greater size, age). Randomization testing showed that only % shrub was significantly related to the configuration of sites, although the result for age suggested a pattern worth investigating further.

Variable	*n*	*r*	*P*
Area	28	0.28	0.380
Shrub	28	0.33	0.250
% shrub	28	0.69	0.010
Distance nearest source	25	0.18	0.740
Distance nearest fragment	25	0.20	0.640
Age	25	0.47	0.050

Box 18-4 Worked example of enhanced MDS: geographic variation and forest bird assemblages.

The data set from Mac Nally (1995) consisted of the maximum abundance (from four seasons) for 102 species of birds for 37 sites in southeastern Australia. A matrix of Bray-Curtis dissimilarities between sites was constructed. No standardization was used because the data were densities of birds, rather than absolute counts. This means that species with high densities will dominate the dissimilarities between sites. A nonmetric MDS in two dimensions, using 20 random starts, resulted in a stress value of 0.14. Using three dimensions, a stress value of 0.08 was achieved from 12 of the 20 random starts, so the three dimensional solution was used. The scaling/ordination plot of the 37 sites in the first two of the three dimensions (Figure 18-5a) showed clear separation of sites dominated by Gippsland Manna Gum and River Red Gum, and to a lesser extent Box-Ironbark. The remaining habitat types (Foothills woodland and Montane forest) could not be easily distinguished from the unclassified sites. If we had no evidence for prior groupings in these data, we might use a minimum spanning tree to further examine relative closeness of sites (Figure 18-5b). The three longest spans would roughly separate the River Red Gum and Gippsland Manna Gum habitats from the rest, with two of the unclassified sites intermediate.

Mac Nally (1995) was able to classify the sites *a priori* into five habitat types so we were able to test the H_0 of no difference between the five habitat types using a single factor ANOSIM procedure. We used the program PRIMER. The global *R* statistic was 0.914 and the probability of obtaining a value this great or greater, based on a randomization test, was less than 0.001. We concluded there were statistically significant differences in bird assemblages between habitats. Pairwise ANOSIM tests were difficult to interpret because there were only four observations in each group which only allowed 35 possible permutations for each pairwise randomization test and thus *P* values were ± approximately 0.029. However, only two of the pairwise comparisons had *R* values less than one, Montane forest versus Foothills woodland and Box-Ironbark versus Foothills woodland.

We also used the nonparametric MANOVA procedure of Anderson (2001) to test the H_0 of no difference between the five habitat types. We used the program NP-MANOVA, kindly supplied by M.J. Anderson from the University of Auckland. The single factor MANOVA test was based on Bray-Curtis dissimilarities between sites and we used 10,000 permutations.

Source	SS	df	MS	*F*	*P*	Possible number of permutations
Habitat	28964.903	4	7241.226	9.619	< 0.001	2.55×10^9
Residual	11292.165	15	752.811			
Total	40257.068	19				

Clearly, we would reject the H_0 and conclude that there is a significant difference across the five habitats in the Bray-Curtis dissimilarities between sites. We then ran pairwise comparisons, based on *t* statistics ($\sqrt{F}$ from nonparametric MANOVA comparing two groups). All comparisons were significant, except Foothills woodland v Montane forest. indicating that this procedure is more powerful than the ANOSIM tests, although Holm's adjustment to the *P* values to control the familywise Type I error rate resulted in no significant differences (all $P = 0.280$).

Comparison	*t*	*P*

Box-Ironbark v Foothills woodland	1.702	0.031
Box-Ironbark v Gippsland Manna Gum	3.227	0.028
Box-Ironbark v Montane forest	2.676	0.028
Box-Ironbark v River Red Gum	3.639	0.028
Foothills woodland v Gippsland Manna Gum	2.954	0.031
Foothills woodland v Montane forest	1.520	0.054
Foothills woodland v River Red Gum	3.550	0.028
Gippsland Manna Gum v Montane forest	3.262	0.029
Gippsland Manna Gum v River Red Gum	3.361	0.028
Montane forest v River Red Gum	4.287	0.030

Box 18-5 Worked example of cluster analysis: genetic structure of a rare plant

Like MDS, hierarchical cluster analysis can be based on any type of dissimilarity matrix. We clustered the data on the eight subpopulations of the rare annual plant (*Clarkia springvillensis*) in California based on Cavalli-Svorza genetic distances between the subpopulations (McCue *et al*. 1996). We used UPGMA and the dendogram is shown in Figure 18-6. The two Gauging Station subpopulations (GS) split off first; these were most different in the NMDS scaling plot based on the same matrix – see Box 18-2. Then the second of the Springville Clarkia Ecological Reserve (SCER) subpopulations grouped with the second and third Bear Creek (BC) subpopulations and the first BC subpopulation grouped with the first and third SCER subpopulations.

Box 18-6 Worked example of cluster analysis: geographic variation and forest bird assemblages

A matrix of Bray-Curtis dissimilarities, based on densities of 102 species of birds, between sites was used to hierarchically structure the 37 sites in southeastern Australia Mac Nally 1995). No standardization was used because the data were densities of birds, rather than absolute counts. The UPGMA clustering procedure produced the dendogram shown in Figure 18-7a, although representing this in polar form (Figure 18-7b) makes presentation a little easier. The Gippsland Manna Gum sites, the River Red Gum sites and the Box-Ironbark sites grouped into clear clusters, whereas the remaining habitat types (Foothills woodland and Montane forest) were not in separate clusters. This interpretation is similar to that from the NMDS on the same matrix of dissimilarities (Box 18-5).

Figure 18-1 PCoA scaling/ordination plot of the 28 sites from Bolger et al. *(1997) based on a Bray-Curtis matrix of dissimilarities between sites, standardized so all sites have maximum abundance of one. The three mainland sites are filled symbols.*

Figure 18-2 Illustration of the links between (a) configuration (scaling/ordination) plot, (b) dissimilarity matrix and (c) Shepard plot in enhanced MDS. S_1, S_2 etc. are objects, e.g. sampling units.

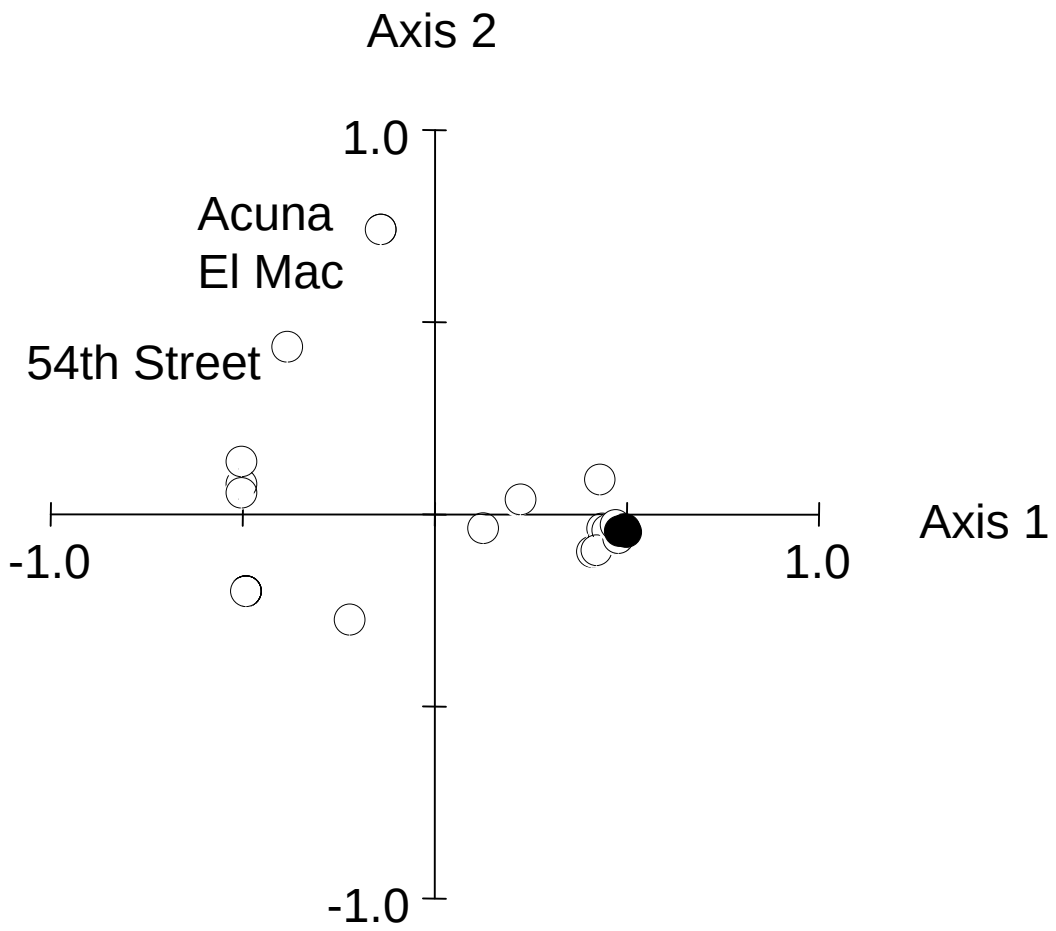
Figure 18-3 (a) NMDS scaling/ordination plot of the eight subpopulations of the plant Clarkia springvillensis based on Cavalli-Svorza genetic distances between subpopulations; from McCue et al. (1996) (b) Shepard plot showing the relationship between Cavalli-Svorza genetic distances between subpopulations and NMDS distances between subpopulations.

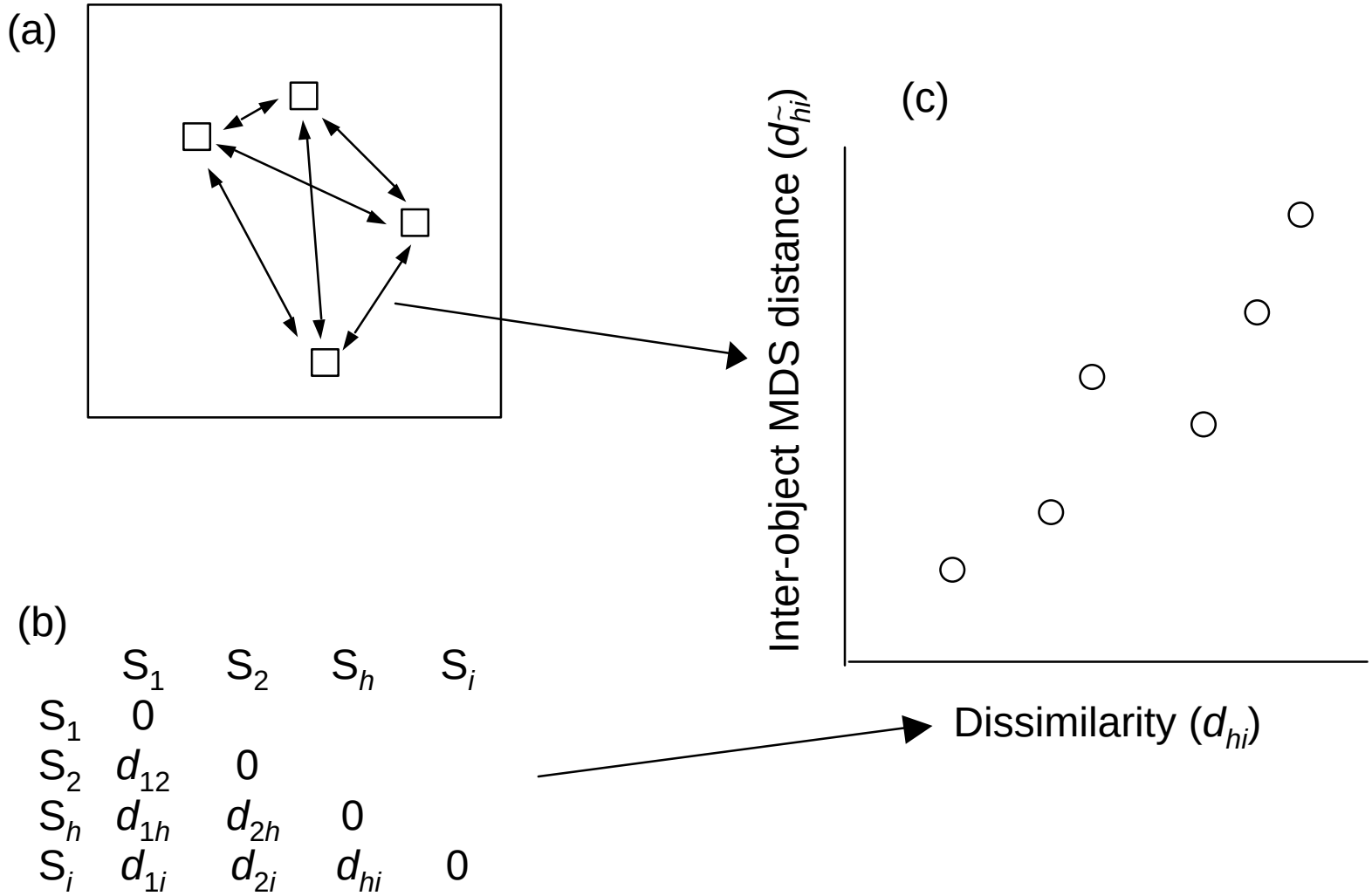
Figure 18-4 NMDS scaling/ordination plot of the 28 sites from Bolger et al. (1997) based on a Bray-Curtis matrix of dissimilarities between sites, standardized so all sites have maximum abundance of one. The three mainland sites are filled symbols.

Figure 18-5 NMDS scaling/ordination plots of the 37 sites from Mac Nally (1995) based on a Bray-Curtis matrix of dissimilarities between sites. In (a), the different habitats are identified by different symbols. In (b), a minimum spanning tree joins all sites with longest spans indicated by ***.***

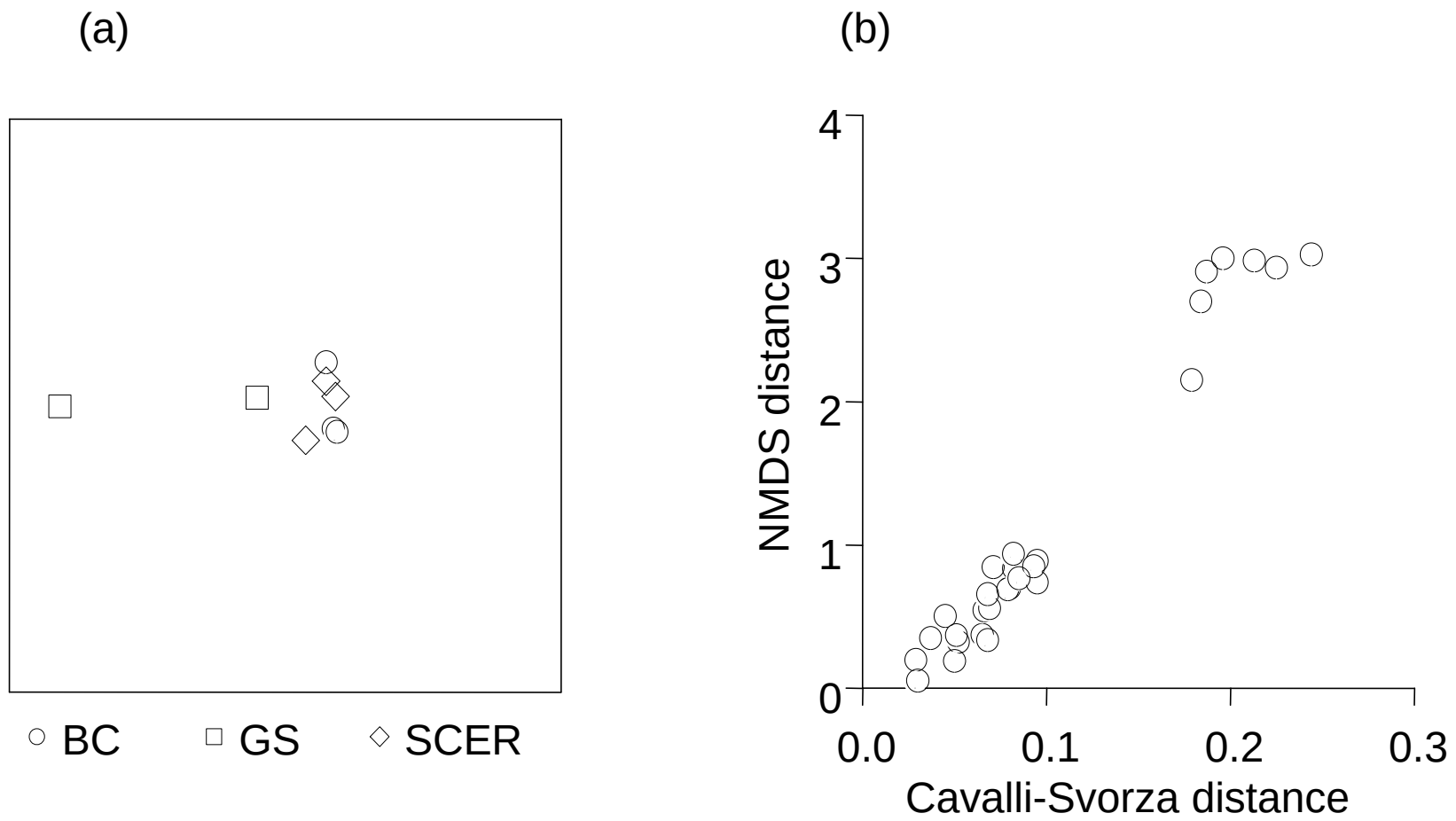
Figure 18-6 Dendogram from hierarchical UPGMA cluster analysis of the eight subpopulations of the plant Clarkia springvillensis based on Cavalli-Svorza genetic distances between subpopulations; from McCue et al. (1996).

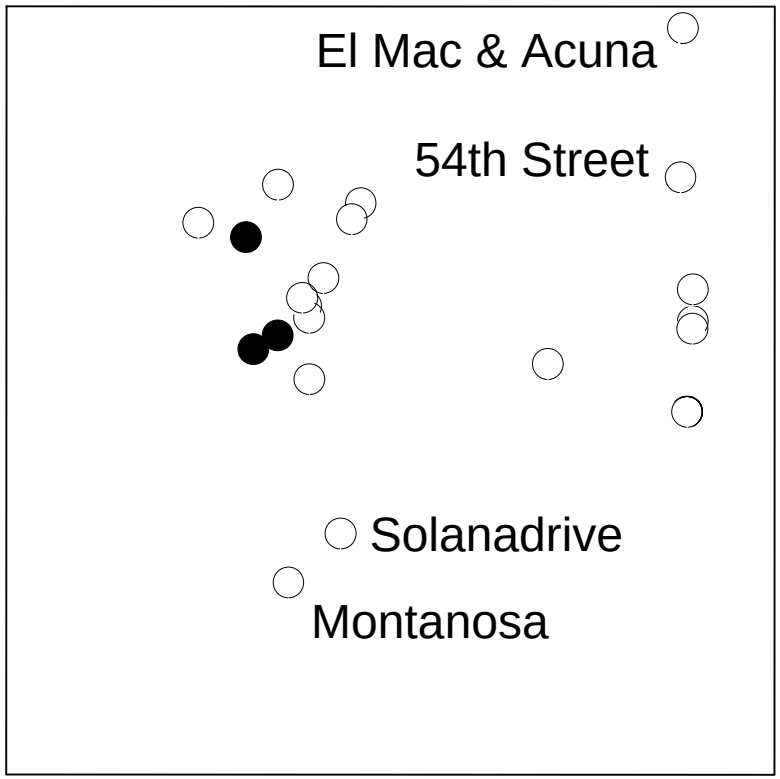
Figure 18-7 Dendograms from hierarchical UPGMA cluster analysis of the 37 sites from Mac Nally (1995) based on a Bray-Curtis matrix of dissimilarities between sites. In (a), the usual dendogram is displayed with clusters identified for Gippsland Manna Gum (Gr 1), Box-Ironbrak (Gr 4) and River Red Gum (Gr 5). In (b), the polar representation of the dendogram is displayed, with site numbers. Gippsland Manna Gum includes sites 2, 3, 4 and 24; Montane forest sites 9, 11, 12, 15; Foothills woodland sites 10, 20, 21, 37; Box-Ironbark sites 25, 33, 34, 36; River Red Gum sites 29,30, 31, 32; remaining sites unclassified.

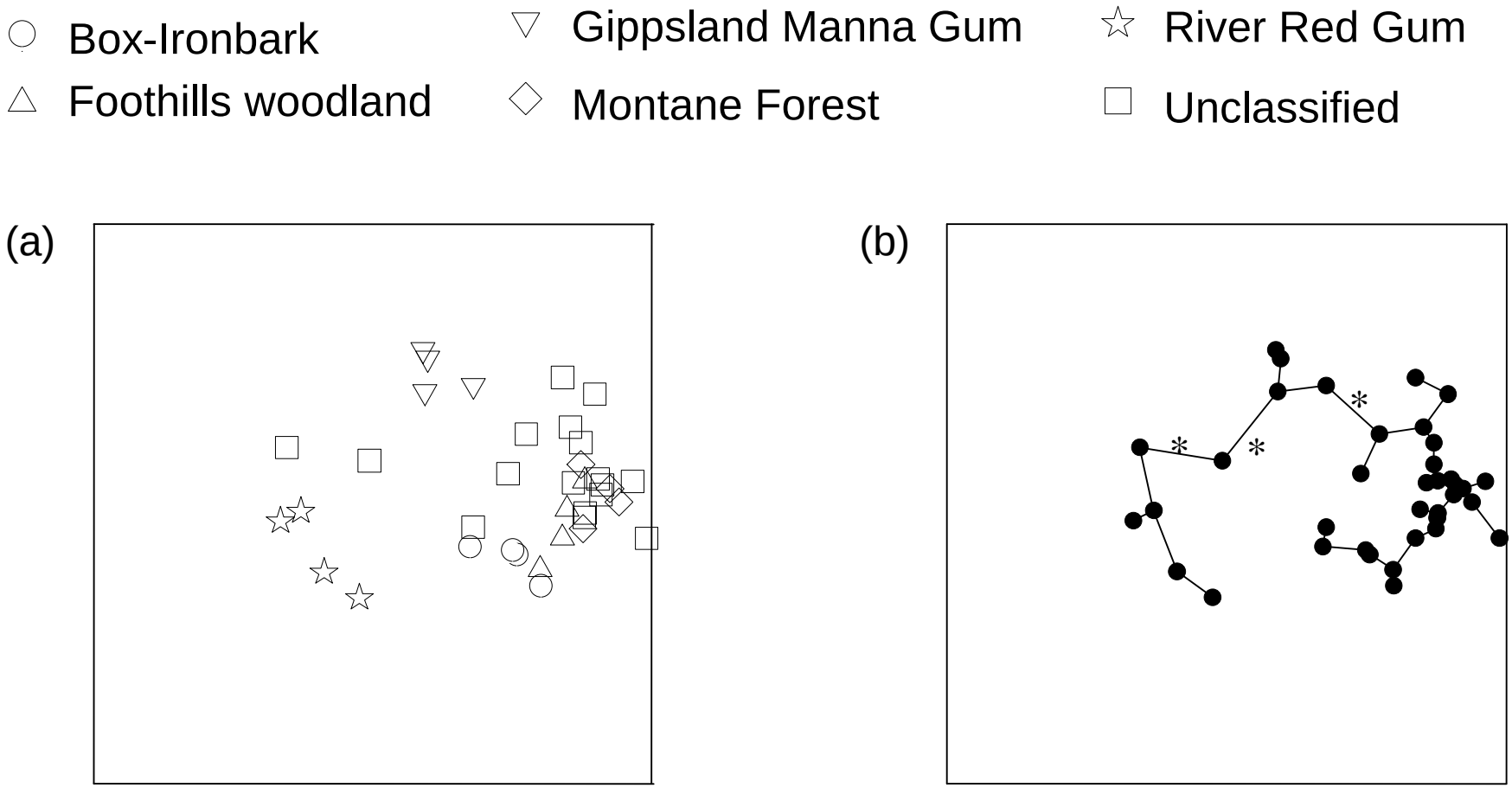

Axis 2
1.0
Acuna
El Mac
54th Street
-1.0
1.0
Axis 1
-1.0

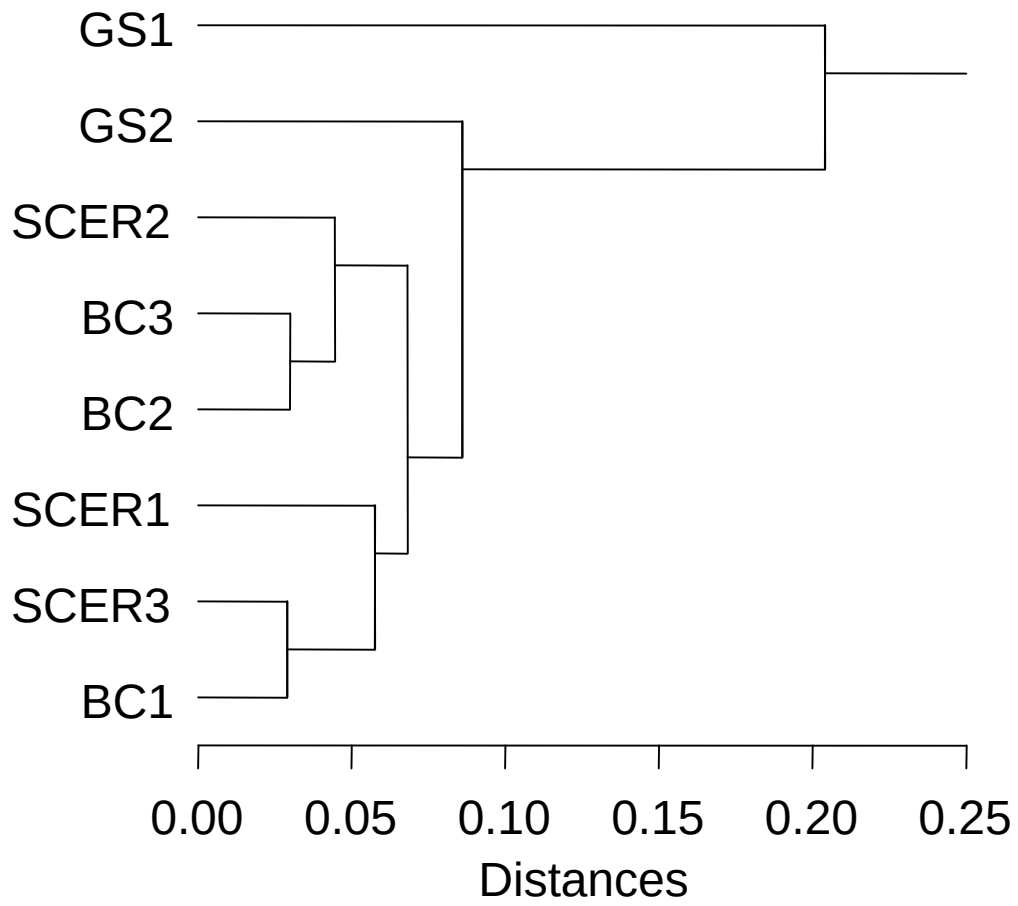


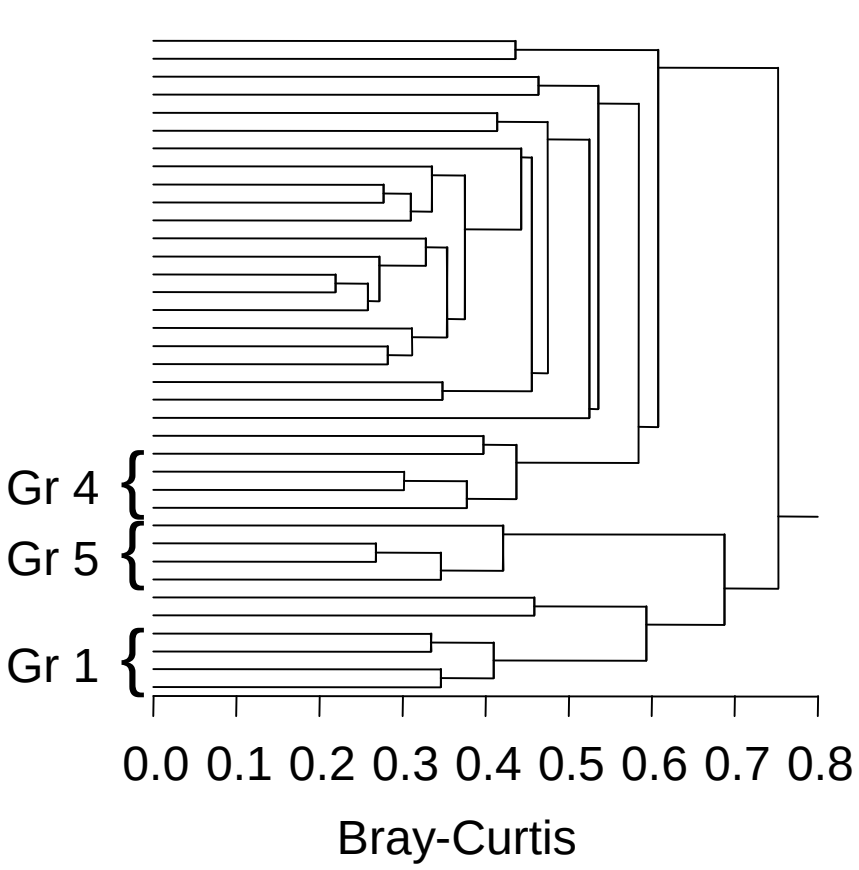
	S_1	S_2	S_h	S_i
S_1	0			
S_2	d_{12}	0		
S_h	d_{1h}	d_{2h}	0	
S_i	d_{1i}	d_{2i}	d_{hi}	0

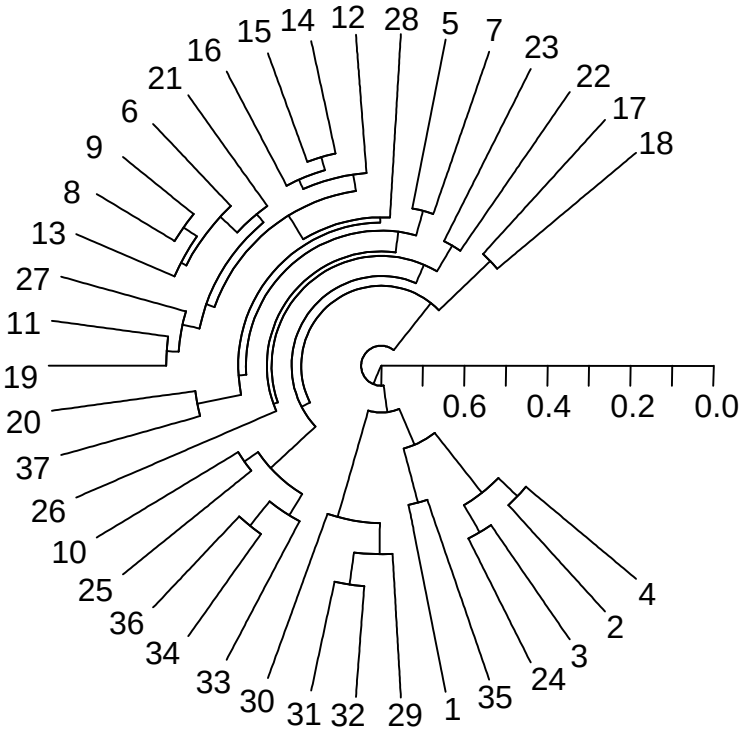

(a)
(b)
BC
GS
SCER
NMDS distance
Cavalli-Svorza distance
0
1
2
3
4
0.0
0.1
0.2
0.3


El Mac & Acuna
54th Street
Solanadrive
Montanosa


Box-Ironbark
Foothills woodland
Gippsland Manna Gum
Montane Forest
River Red Gum
Unclassified
(a)
(b)


GS1
GS2
SCER2
BC3
BC2
SCER1
SCER3
BC1
0.00
0.05
0.10
0.15
0.20
0.25
Distances


Gr 4
Gr 5
Gr 1
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
Bray-Curtis


15 14 12 28 5 7 23 22 17 18
16 21 6 9 8 13 27 11 19 20 37 26 10 25 36 34 33 30 31 32 29 1 35 24 3 2 4
0.6 0.4 0.2 0.0