Manuscript of – Legendre, P. and H. J. B. Birks. 2012. From classical to canonical ordination. Chapter 8, pp. 201-248 in: *Tracking Environmental Change using Lake Sediments, Volume 5: Data handling and numerical techniques*. H. J. B. Birks, A. F. Lotter, S. Juggins and J. P. Smol [eds.]. Springer, Dordrecht, The Netherlands. xi + 716 pp.

8. FROM CLASSICAL TO CANONICAL ORDINATION

PIERRE LEGENDRE (Pierre.Legendre@umontreal.ca)
Département de sciences biologiques,
Université de Montréal,
C.P. 6128, succursale Centre-ville, Montréal,
Québec H3C 3J7, Canada

and

H. JOHN B. BIRKS (John.Birks@bio.uib.no)
Department of Biology and Bjerknes Centre for Climate Research,
University of Bergen,
PO Box 7803,
N-5020 Bergen, Norway

Environmental Change Research Centre,
Pearson Building, Gower Street
University College London,
London, WC1E 6BT, UK

and

School of Geography and the Environment,
University of Oxford,
Oxford, OX1 3QY, UK

Keywords*:* canonical correspondence analysis, correspondence analysis, non-metric multidimensional scaling, ordination, PCNM analysis, principal component analysis, principal coordinate analysis, redundancy analysis, variation partitioning, principal coordinates of neighbour matrices

Introduction

To ordinate is to arrange objects in some order (Goodall 1954). Ordination procedures are well-known to ecologists who wish to represent and summarise their observations along one, two, or a few axes. The most simple case is the ordination of sites along a single variable representing an environmental gradient (e.g., water temperature, pH), or a sampling variable such as depth along a sediment core or along the estimated ages of levels in a sediment core. Ordination diagrams are simply scatter plots of the objects (e.g., core levels) on two or sometimes three axes according to the values taken by the objects along the variables comprising the axes.

When the data are multivariate, the problem is to either choose two pertinent variables for plotting the observations, or to construct synthetic variables that represent, in some optimal mathematical way, the set of variables under study; these synthetic variables may then be used as the major axes for the ordination. The data matrix subjected to analysis may contain a set of environmental variables, or the multi-species composition of the assemblage under study. In such cases, we will say that we are performing an ordination in a space of reduced dimensionality, or an ordination in reduced space, since the original data-set has many more dimensions (variables) than the ordination graph we want to produce.

This chapter describes the choices that have to be made in order to obtain a meaningful and useful ordination diagram. It will also show how the methods of canonical ordination, which are widely used to relate species to environmental data in palaeolimnology, are extensions within the framework of regression modelling of two classical ordination methods. Some forms of ordination analysis, classical or canonical, are widely used by palaeolimnologists as tools in the handling, summarisation, and interpretation of palaeolimnological data, either modern assemblages or core fossil assemblages. The various types of use of ordination analysis in palaeolimnology are summarised in Table 1. No attempt is made here to provide a comprehensive review of palaeolimnological applications of ordination methods. Emphasis is placed instead on basic concepts and the critical methodological questions that arise in the use of ordination methods in palaeolimnology. Birks (2008, 2010) provides a short overview of the range of ordination methods currently available and of the general use and value of ordination techniques in ecology and palaeoecology. Borcard et al. (2011) discuss classical (unconstrained) and canonical (constrained) ordinations and their implementation with R.

Basic concepts in simple ordination

The simple ordination methods mostly used by (palaeo)ecologists and (palaeo)limnologists are principal component analysis (PCA), correspondence analysis (CA) and its relative, detrended correspondence analysis (DCA), principal coordinate analysis (PCoA), and non-metric multidimensional scaling (NMDS) (Prentice 1980, 1986). These methods will be reviewed here in a geometric framework. They mostly differ in the types of distances among objects that they attempt to preserve in the ordination.

Simple ordination is used in palaeolimnology to address two main types of questions. (1) In a study of sediment cores, ordination is used to identify the main gradients in the species assemblage data, which are multivariate by nature, and to interpret these gradients using species loadings on the ordination axes (see Birks: Chapter 9 this volume). Ordinations are also used as graphical templates to draw groups of sampling units obtained by clustering, as well as trajectories of the multivariate species data through time to estimate the magnitude and rates of change in species assemblage composition (Birks and Gordon 1985; Jacobson and Grimm 1986; Birks 1992; Birks: Chapter 9 this volume). (2) Ordination of modern objects from various locations is also used as a basis on which fossil objects can be projected as passive objects for comparison between modern and fossil assemblages (Lamb 1984; Birks and Gordon 1985; Birks 1992, Chapter 9 this volume).

Starting with a data-set, several choices have to be made before obtaining an ordination (Table 2). These choices will be described in some detail because a good understanding of their implications is likely to produce more informative and useful ordination diagrams. Users of ordination methods should not let themselves be guided blindly by the implicit choices that are inherent to some methods or computer programs. The critical decisions to be made are the following:

- Do the data (environmental or assemblage data) need to be transformed prior to ordination analysis?

- Which distance measure should be preserved by the ordination method?
- Should a metric or non-metric ordination method be used?
- How many axes are required?

These decisions will now be discussed in some detail.

Transformation of physical data

Physical, chemical, or geological variables are often used as explanatory variables in palaeolimnological studies. They may also be used directly to obtain ordinations of the objects or sites on the basis of these variables (Table 1). Three problems may require pre-processing of the data before ordination: (a) if the distributions of the data along the variables are not symmetrical, skewness may need to be reduced; (b) if the variables are not all expressed in the same physical units, they need to be transformed to eliminate their physical units; (c) multistate qualitative variables (e.g., rare, common, abundant) may require, in some cases, transformation into dummy variables prior to ordination.

a) An ordination in which some of the points are clumped in a big mass while other points are stretched across the diagram is not very useful or informative. It is better to have the points scattered in a fairly homogeneous fashion across the diagram, with perhaps some clumping in the centre of the diagram, or in some areas of higher density if the data are clumped; the latter case may suggest that a cluster analysis might produce a more interesting and useful multivariate description of the data (see Legendre and Birks: Chapter 6 this volume).

The data should be initially examined using univariate methods, such as computing skewness statistics, or drawing frequency histograms (see Juggins and Telford: Chapter 4 this volume). Depending on the type of asymmetry found, various transformations can be applied, such as square root, double square root, or log transformation. General methods, such as the Box-Cox transformations, are available to find automatically the most efficient normalising transformation; see Sokal and Rohlf (1995) or Legendre and Legendre (1998). These are often referred to as *normalizing transformations* because removing the asymmetry is an important step towards obtaining normally-distributed data. We emphasise, however, that the objective prior to ordination is not to obtain a multinormal distribution of the data, but simply to reduce the asymmetry of the distributions. Tests of normality may be useful to screen the data and identify the variables whose distributions should be examined more closely in order to find, if possible, a skewness-reducing transformation (see Juggins and Telford: Chapter 4 this volume).

Scientists often worry about transforming variables. Is it permissible? The original physical unit in which an environmental variable is measured imposes a scale to the data that is as unlikely to be related to the response of the species to this variable as any other scale that we may impose by applying a nonlinear transformation to the data. Physiological studies would be required to determine what the most appropriate transformation is, in order to relate a physical variable to the response of the species. So, short of having such information available to them, users of ordination methods are left with statistical criteria only, such as skewness of the distributions, to decide on the transformation of physical variables.

b) In most cases, physical variables are not expressed in the same physical units; some may be in cm, others in μg L^{-1}, in °C, or in pH units. Such variables need to be transformed to eliminate the physical dimensions before being used together to produce an ordination. Note that log-transformed data are dimensionless because logarithms are exponents of a base and exponents are dimensionless. There are two main methods for eliminating physical dimensions: standardisation and ranging. They both eliminate the physical units by dividing the original data by a value possessing the same physical units.

- Standardise variable **y** to **z**: $z_i = \frac{y_i - \bar{y}}{s_y}$ (1)

  where y_i is the original value of variable **y** for object i, $\bar{y}$ is the mean value of **y**, and s_y is the estimated standard deviation of **y**. z_i is the standardised value of variable **y** for object i. Variable standardisation is available in the 'decostand' function of the VEGAN R-language package (method = "standardize").

- Relative-scale variables: range variable **y** to **y**' using equation $y_i' = y_i / y_{max}$ (2)

  where y_i' is the ranged value of **y** for object i and y_{max} is the maximum value of **y** in the whole data table. This form of ranging is used for relative-scale variables, where 'zero' means the absence of the characteristic of interest. This transformation is available in the 'decostand' function of the VEGAN R-language package (method = "max").

- Interval-scale variables: range variable **y** to **y**' using equation $y_i' = \frac{y_i - y_{min}}{y_{max} - y_{min}}$ (3)

  where y_i' is the ranged value of **y** for object i, whereas y_{min} and y_{max} are, respectively, the minimum and maximum values of **y** in the whole data table. This form of ranging is used for interval-scale variables, in which the value 'zero' is chosen arbitrarily and whose range may include negative values. Temperatures in °C are an example of an interval-scale variable. This transformation is available in the 'decostand' function of the VEGAN R-language package (method = "range").

Variables may also be standardised in order to bring their variances to unity. It is preferable to apply skewness-reducing transformations before standardising the data. If the opposite is done, standardisation would produce negative values which are incompatible with square root, log, or Box-Cox transformations. Ranging, which brings all values of a variable into the interval [0,1], may be used before or after applying a skewness-reducing transformation.

c) Multistate qualitative variables may be handled in different ways. If the ordination is to be obtained through a method requiring the prior calculation of a distance matrix (PCoA, NMDS), resemblance coefficients are available that are capable of handling mixtures of quantitative and qualitative variables, as discussed in the section below on Choice of an appropriate distance function and in Simpson (Chapter 13 this volume). If, on the other hand, the ordination is to be obtained through a method that will implicitly preserve the Euclidean distance among objects (PCA, redundancy analysis (RDA)), the qualitative data must be transformed in some way prior to being subjected to the ordination method because a qualitative variable is not a *metric* or *measurement* variable; in other words, the distance between states 1 and 3 of a qualitative variable is not twice as large as the distance between states 1 and 2. Variables from which Euclidean distances are calculated must be metric (or quantitative). The transformation can be done in one of two ways:

- A qualitative variable possessing p states can be recoded into p binary (0-1) variables, called *dummy variables*, using one dummy variable for each state of the qualitative variable. The coding method is described in Legendre and Legendre (1998, Chapter 1). Dummy variables can be used in PCA or RDA only if the program provides a possibility for weighting the variables. Indeed, if the variables are standardised or ranged prior to the ordination, a qualitative variable recoded into *p* dummy variables occupies *p* dimensions in the full-dimensional representation of the data. Each dummy variable should be downweighted to have a weight of 1/*p* in the analysis while the other quantitative variables have a weight of 1. The program CANOCO (ter Braak and Šmilauer 2002) offers the possibility of specifying weights for variables in PCA or RDA.

- Redundancy analysis (RDA) or canonical correspondence analysis (CCA) can be used to find a transformation of a qualitative multistate variable into a quantitative variable which is optimal with respect to a table of assemblage composition data (Legendre and Legendre, 1998: 597). This is done as follows. Recode the qualitative variable into dummy variables as in the previous paragraph. Remove one of the dummy variables because, with all *p* dummy variables, the variance-covariance matrix of the dummy variables is singular and cannot be inverted; this is an obligatory step for explanatory variables in multiple regression and canonical analysis. For RDA or CCA, use the table of species composition data as the response matrix and the table of dummy variables as the explanatory matrix. If the first canonical ordination axis explains most of the canonical variance, it can be used in further analyses as a quantitative representation of the original qualitative variable. [Note: in the program CANOCO, the last of a set of dummy variables is automatically removed from the calculations. In the same way, the last state of a ‘factor’ variable is removed from the calculations in the ‘rda’ and ‘cca’ functions of the VEGAN R-language package, but the centroids of all states are drawn in the biplot.]

Transformation of assemblage composition data

Assemblage composition data (species abundances) for short gradients, which contain relatively few zeros, can be ordinated by PCA or RDA: in that case the Euclidean distance is a meaningful measure of the ecological distance among the observations. These variables may, however, have asymmetric distributions because species tend to have exponential growth when conditions are favourable. This well-known fact has been embedded in the theory of species-abundance models; see He and Legendre (1996, 2002) for a synthetic view of these models. To reduce the asymmetry of the distributions, species abundance **y** may be transformed to **y**' by taking the square root or the fourth root (which is equivalent to taking the square root twice), or by using a log transformation:

$$y' = y^{0.5} \qquad \text{or} \qquad y' = y^{0.25} \qquad \text{or} \qquad y' = \log(y + c) \tag{4}$$

where y is the species abundance and c is a constant. Usually, $c = 1$ in species log transformations, so that an abundance $y = 0$ is transformed into $y' = \log(0 + 1) = 0$ for any logarithmic base. Michael Palmer (http://www.okstate.edu/artsci/botany/ordinate/) does not recommend this transformation for absolute biomass data because it gives different values depending on the mass units (e.g., g or kg) used to record biomass. Another transformation that reduces the asymmetry of heavily skewed abundance data is the one proposed by Anderson et al. (2006). The abundance data y_{ij} are transformed to an exponential scale that makes allowance for zeros: y'_{ij} =

log$_{10}$(yij) + 1 when $y_{ij} > 0$ or $y'_{ij} = 0$ when $y_{ij} = 0$. Hence, for $y_{ij} = \{0, 1, 10, 100, 1000\}$, the transformed values are $\{0, 1, 2, 3, 4\}$. This transformation is available in the ***decostand()*** function of the VEGAN package (method = "log").

Community composition data sampled along long ecological gradients typically contain many zero values because species are known to generally have unimodal responses along environmental gradients (ter Braak and Prentice 1988). The proportion of zeros is greater when the sampling has crossed a long environmental gradient. This is because species have optimal niche conditions, where they are found in greater abundances along environmental variables (see juggins and Birks: Chapter 12 this volume). The optimum for a species along an environmental variable corresponds to the centre of its theoretical Hutchinsonian niche along that factor. These propositions are discussed in most texts of community ecology and, in particular, in Whittaker (1967) and ter Braak (1987c). Because ordination methods use a distance function as their metric to position the objects with respect to one another in ordination space, it is important to make sure that the chosen distance is meaningful for the objects under study. Choosing an appropriate distance measure means trying to model the relationships among the sites appropriately for the assemblage composition data at hand. The choice of a distance measure is an ecological, not a statistical decision.

An example presented in Legendre and Legendre (1998, p. 278) shows that the Euclidean distance function may produce misleading results when applied to assemblage composition data. Alternative (dis)similarity functions described in the next section, which were specifically designed for assemblage composition data, do not have this drawback. In some cases, distance measures that are appropriate for assemblage composition data can be obtained by a two-step procedure: first, transform the species abundance data in some appropriate way, as described below; second, compute the Euclidean distance among the sites using the transformed data (Figure 1). This also means that assemblage composition data transformed in these ways can be directly used to compute ordinations by the Euclidean-based methods of PCA or RDA; this approach is called transformation-based PCA (tb-PCA) or transformation-based RDA (tb-RDA). The transformed data matrices can also be used in *K*-means partitioning, which is also a Euclidean-based method (see Legendre and Birks: Chapter 6 this volume). Legendre and Gallagher (2001) have shown that the following transformations can be used in that context (some of these transformations have been in use in community ecology and palaeoecology for a long time, e.g., by Noy-Meir et al. (1967) and by Prentice (1980)).

1) Transform the species abundances from each object (sampling unit) into a vector of length 1, using the equation:

$$y'_{ij} = y_{ij} \Big/ \sqrt{\sum_{j=1}^{p} y_{ij}^2} \qquad (5)$$

where y_{ij} is the abundance of species j in object i. This equation, called the ‘chord transformation’ in Legendre and Gallagher (2001), is one of the transformations available in the program CANOCO (Centring and standardisation for “samples”: *Standardise by norm*) and in the ‘decostand’ function of the VEGAN R-language package (method = “normalize”). If we compute the Euclidean distance

$$D_{\text{Euclidean}}(\mathbf{x'}_1, \mathbf{x'}_2) = \sqrt{\sum_{j=1}^{p} (y'_{1j} - y'_{2j})^2} \tag{6}$$

between two rows ($\mathbf{x'}_1$, $\mathbf{x'}_2$) of the transformed data table, the resulting value is identical to the chord distance (equation 18) that could be computed between the rows of the original (untransformed) species abundance data table (Figure 1). The interest of this transformation is that the chord distance, proposed by Orlóci (1967) and Cavalli-Sforza and Edwards (1967), is one of the distances recommended for species abundance data. Its value is maximum and equal to √2 when two objects have no species in common. As a consequence, after the chord transformation, the assemblage composition data are suitable for PCA or RDA which are methods preserving the Euclidean distance among the objects.

2) In the same vein, if the data [y_{ij}] are subjected to the 'chi-square distance transformation' as follows:

$$y'_{ij} = \sqrt{y_{++}} \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \tag{7}$$

where y_{i+} is the sum of the row (object) values, y_{+j} is the sum of the column (species) values, and y_{++} is the sum of values of the whole data table, then Euclidean distances computed among the rows of the transformed data table [y'_{ij}] are equal to chi-square distances (equation 19) among the rows of the original, untransformed data-table. The chi-square distance, preserved in correspondence analysis, is another distance often applied to species abundance data. Its advantage or disadvantage, depending upon the circumstances, is that it gives higher weight to the rare than to the common species. The chi-square distance transformation is available in the 'decostand' function of the VEGAN R-language package (method = "chi.square").

3) The data can be transformed into profiles of relative species abundances through the equation:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}} \tag{8}$$

which is a widespread method of data standardisation, prior to analysis, especially when the sampling units are not all of the same size as is commonly the case in palaeolimnology. Data transformed in that way are called *compositional data*. In palaeolimnology and community ecology, the species assemblage is considered to represent the response of the community to environmental, historical, or other types of forcing; the variation of any single species has no clear interpretation. Compositional data are used because ecologists and palaeoecologists feel that the vectors of relative proportions of species can lead to meaningful interpretations. Many fossil or recent assemblage data-sets are presented as profiles of relative abundances, for example, in palynology and palaeolimnology, or as percentages if the values y'_{ij} are multiplied by 100. Computing Euclidean distances among rows (objects) of a data-table transformed in this way produces 'distances among species profiles' (equation 20). The transformation to profiles of relative abundances is available in the 'decostand' function of the VEGAN R-language package (method =

"total"). Statistical criteria investigated by Legendre and Gallagher (2001) show that this is not the best transformation; the Hellinger transformation (next paragraph) is preferable. Log-ratio analysis has been proposed as a way of analysing compositional data (Aitchison 1986). This method is, however, only appropriate for data that do not contain many zeros (ter Braak and Šmilauer 2002).

4) A modification of the species profile transformation is the Hellinger transformation:

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \qquad (9)$$

Computing Euclidean distances among rows (objects) of a data table transformed in this way produces a matrix of Hellinger distances among sites (equation 21). The Hellinger distance, described in more detail below, is a measure recommended for clustering or ordination of species abundance data (Prentice 1980; Rao 1995). It has good statistical properties as assessed by the criteria investigated by Legendre and Gallagher (2001). The Hellinger transformation is available in the 'decostand' function of the VEGAN R-language package (method = "hellinger").

Before using these transformations, one may apply a square root or log transformation to the species abundances in order to reduce the asymmetry of the species distributions (Table 2). The transformations described above can also be applied to presence-absence data. The chord and Hellinger transformations appear to be the best for general use. The chi-square distance transformation is interesting when one wants to give more weight to the rare species; this is the case when the rare species are considered to be good indicators of special ecological conditions. We will come back to the use of these transformations in the following sections. Prior to these transformations, any of the standardisations investigated by Noy-Meir et al. (1975), Prentice (1980), and Faith et al. (1987) may also be used if the study justifies it: species adjusted to equal maximum abundances or equal standard deviations, sites standardised to equal totals, or both.

Choice of an appropriate distance function

Most statistical and numerical analyses assume some form of distance relationship among the observations. Univariate and multivariate analyses of variance and covariance, for instance, assume that the Euclidean distance is the appropriate way of describing the relationships among objects; likewise for methods of multivariate analysis such as *K*-means partitioning and PCA (see Legendre and Birks: Chapter 6 this volume). It is the responsibility of the scientist doing the analyses to either make sure that this assumption is met by the data, or to model explicitly relationships of other forms among the objects by computing particular distance functions and using them in appropriate methods of data analysis.

Many similarity or distance functions have been used by ecologists; they are reviewed by Legendre and Legendre (1998: Chapter 7), Borcard et al. (2011: Chapter 3) and other authors. We will only mention here those that are most commonly used in the ecological, palaeoecological, and palaeolimnological literature.

1) The Euclidean distance (equation 6) is certainly the most widely used coefficient to analyse tables of physical descriptors, although it is not always the most appropriate. This is the coefficient preserved by PCA and RDA among the rows of the data matrix (objects), so that if the Euclidean distance is considered appropriate to the data, these methods can be applied directly to the data matrix, perhaps after one of the transformations described above, to obtain a meaningful ordination.

2) For physical or chemical data, an alternative to the Euclidean distance is to compute the Gower (1971) coefficient of similarity, followed by a transformation of the similarities to distances. The Gower coefficient is particularly important when one is analysing a table containing a mixture of quantitative and qualitative variables. In this coefficient, the overall similarity is the mean of the similarities computed for each descriptor *j* (see Simpson: Chapter 13 this volume). Each descriptor is treated according to its own type. The partial similarity (s_j) between objects $\mathbf{x}_1$ and $\mathbf{x}_2$ for a quantitative descriptor *j* is computed as follows:

$$s_j(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j} \tag{10}$$

where R_j is the range of the values of descriptor *j* across all objects in the study. The partial similarity s_j is a value between 0 (completely dissimilar) and 1 (completely similar). For a qualitative variable *j*, $s_j = 1$ if objects $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same state of the variable and $s_j = 0$ if they do not. The Gower similarity between $\mathbf{x}_1$ and $\mathbf{x}_2$ is obtained from the equation:

$$S(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{p} s_j(\mathbf{x}_1, \mathbf{x}_2) / p \tag{11}$$

where *p* is the number of variables. The variables may receive different weights in this coefficient; see Legendre and Legendre (1998: 259) for details. See also the note at the end of this section about implementations in R.

For presence-absence of physical descriptors, one may use the simple matching coefficient:

$$S(\mathbf{x}_1, \mathbf{x}_2) = \frac{a + d}{a + b + c + d} = \frac{a + d}{p} \tag{12}$$

where *a* is the number of descriptors for which the two objects are coded 1, *d* is the number of descriptors for which the two objects are coded 0, whereas *b* and *c* are the numbers of descriptors for which the two objects are coded differently. *p* is the total number of physical descriptors in the table.

There are different ways of transforming similarities (*S*) into distances (*D*). The most commonly used equations are:

$$D(\mathbf{x}_1, \mathbf{x}_2) = 1 - S(\mathbf{x}_1, \mathbf{x}_2) \tag{13}$$

and $$D(\mathrm{x}_1,\mathrm{x}_2) = \sqrt{1 - S(\mathrm{x}_1,\mathrm{x}_2)} \qquad (14)$$

For the coefficients described in equations 11, 12, 15, 16, 17, equation 14 is preferable for transformation prior to ordination because the distances so obtained produce a fully Euclidean representation of the objects in the ordination space, except possibly in the presence of missing values; equation 13 does not guarantee such a representation (Legendre and Legendre 1998, Table 7.2). The concept of Euclidean representation of a distance matrix is explained below in the section on Euclidean or Cartesian space, Euclidean representation. Equation 14 is used for transformation of all binary coefficients computed by the 'dist.binary' function of the ADE4 R-language package.

3) For species presence-absence data,

1. the Jaccard coefficient:

$$S(\mathbf{x}_1,\mathbf{x}_2) = \frac{a}{a+b+c} \qquad (15)$$

2. and the Sørensen coefficient of similarity:

$$S(\mathbf{x}_1,\mathbf{x}_2) = \frac{2a}{2a+b+c} \qquad (16)$$

are widely used. In these coefficients, *a* is the number of species that the two objects have in common, *b* is the number of species found at site or sample 1 but not at site or sample 2, and *c* is the number of species found at site or sample 2 but not at site or sample 1. In order to obtain a fully Euclidean representation of the objects in the ordination space, these similarities should be transformed into distances using equation 14.

3. The Ochiai (1957) coefficient:

$$S(\mathbf{x}_1,\mathbf{x}_2) = \frac{a}{\sqrt{(a+b)(a+c)}} \qquad (17)$$

deserves closer attention on the part of palaeoecologists since it is monotonically related to the binary form of the widely used chord and Hellinger distances described below (equations 18 and 21).

For ordination analysis, the three similarity coefficients described above (equations 15, 16, and 17) can be transformed into Euclidean-embeddable distances using the transformation $D(\mathbf{x}_1,\mathbf{x}_2) = \surd(1 - S(\mathbf{x}_1,\mathbf{x}_2))$ (equation 14). These distances will not produce negative eigenvalues in principal coordinate analysis and will thus be entirely represented in Euclidean space.

An interesting similarity coefficient among sites, applicable to presence-absence data, has been proposed by the palaeontologists Raup and Crick (1979): the coefficient is the probability of the data under the hypothesis of no association between objects. The number of species in common in two sites, *a*, is tested for significance under the null hypothesis H_0 that there is no association between sites $\mathbf{x}_1$ and $\mathbf{x}_2$ because each site in a region (or each level in a core) receives a random subset of species from the regional pool (or the whole sediment core). The association between objects, estimated by *a*, is tested using permutations. The probability (p) that the data conform to the null hypothesis is used as a measure of distance, or (1 – p) as a measure of similarity. The permutation procedure of Raup and Crick (1979) was redescribed by Vellend (2004). Legendre and Legendre (2012, coefficient S_{27}) describe two different permutational procedures that can be used to test the significance of the number of species in common between two sites (i.e. the statistic *a*). These procedures correspond to different null hypotheses. Birks (1989) discusses the application of this and other probabilistic similarity measures in palaeoecology.

4) Several coefficients have been described by ecologists for the analysis of quantitative assemblage composition data. The property that these coefficients share is that the absence of any number of species from the two objects under comparison does not change the value of the coefficient. This property avoids producing high similarities, or small distances, between objects from which many species are absent. The Euclidean distance function, in particular, is not appropriate for assemblage composition data obtained from long environmental gradients because the data table then contains many zeros, and the objects that have many zeros in common have small Euclidean distance values; this is considered to be an inappropriate answer in most ecological and palaeoecological problems. This question is discussed at length in many texts of quantitative community ecology. The coefficients most widely used by ecologists for species abundance data tables are:

1. The chord distance, occasionally called the *cosine-θ distance*:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p}\left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}}\right)^2} = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)} \qquad (18)$$

which consists of subjecting the species data to the chord transformation (equation 5) followed by calculation of the Euclidean distance (equation 6). The chord distance is closely related to the Hellinger distance (equation 20).

2. The chi-square distance:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{y_{++}}\sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}}\left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right)^2} \qquad (19)$$

where y_{i+} is the sum of the frequencies in row *i*, y_{+j} is the sum of the frequencies in column *j*, and y_{++} is the sum of all frequencies in the data table. It is equivalent to subjecting the species data to the chi-square distance transformation (equation 7) followed by calculation of the Euclidean distance (equation 6).

3. The distance between species profiles:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad (20)$$

is equivalent to subjecting the species data to the transformation to profiles of relative abundances (equation 8) followed by calculation of the Euclidean distance (equation 6).

4. The Hellinger distance (Rao, 1995), occasionally called the *chord distance* (Prentice 1980) although it differs from equation 18:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2} \qquad (21)$$

It is equivalent to subjecting the species data to the Hellinger transformation (equation 9) followed by calculation of the Euclidean distance (equation 6). This equation is the chord distance computed on square-root transformed frequencies. In the Hellinger distance, the relative species abundances ("compositional data", used directly in equation 20) are square-root transformed in order to lower the importance of the most abundant species, which may grow exponentially when they encounter favourable conditions. This coefficient thus increases the importance given to the less abundant species (Prentice 1980). The chord (equation 18) and Hellinger (equation 21) functions produce distances in the range [0, √2]. For presence-absence data, they are both equal to $\sqrt{2}\sqrt{1 - \frac{a}{\sqrt{(a+b)\,(a+c)}}}$ where $\frac{a}{\sqrt{(a+b)\,(a+c)}}$ is the Ochiai (1957) similarity coefficient for binary data, described above.

5. A coefficient first described by Steinhaus (*in* Motyka 1947) and rediscovered by other authors, such as Odum (1950) and Bray and Curtis (1957), is:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} \qquad (22)$$

This coefficient has excellent descriptive properties for community composition data (Hajdu, 1981; Gower and Legendre, 1986). Taking the square root of this distance will avoid negative eigenvalues and complex principal axes in principal coordinate analysis. A particular form of this coefficient, for data transformed into percentages by sites (y'_{ij} of equation 8 multiplied by 100), has been described by Renkonen (1938). When presence-absence data are used in equation 22, the resulting coefficient is the one-complement of the Sørensen coefficient of similarity (equation 16) computed over the same data (i.e.,, $D_{(\text{eq. }22)} = 1 - S_{(\text{eq. }16)}$).

6. Whittaker’s (1952) index of association:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}\sum_{j=1}^{p}\left|\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right| \tag{23}$$

7. Clark’s (1952) coefficient of divergence:

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{1}{p}\sum_{j=1}^{p}\left(\frac{y_{1j}-y_{2j}}{y_{1j}+y_{2j}}\right)^2} \tag{24}$$

is a form of the Canberra metric (Lance and Williams 1967) rescaled to the [0, 1] range.

Most of the distances described in this section can be computed using the R-language functions ‘dist’ (STATS package), ‘vegdist’ (VEGAN), ‘dist.binary’ (ADE4), ‘gowdis’ (FD) and ‘daisy’ (CLUSTER); see footnote of Table 3 for references. This statement calls for some remarks. (1) These libraries do not all produce the same results for the binary Jaccard coefficient: ‘dist’ and ‘vegdist’ use the transformation $D = (1 - S)$ (equation 13) whereas ‘dist.binary’ uses $D = \sqrt{1-S}$ (equation 14) to transform similarities into distances. The latter guarantees that a fully Euclidean representation, without negative eigenvalues and complex eigenvectors, will result from principal coordinate analysis. (2) The chord, chi-square and Hellinger distances are not obtained directly but after two calculation steps: transformation of the data (equations 5, 7, 9) followed by calculation of the Euclidean distance (equation 6). (3) Several functions propose the Gower distance: ‘vegdist’ (VEGAN), ‘daisy’ (CLUSTER), and ‘gowdis’ (FD); see footnote of Table 3 for references. The latter is the only function that can handle missing values and variables of all precision levels, including multistate qualitative variables (“factors” in R), and allows users to give different weights to the variables involved in a calculation.

Euclidean or Cartesian space, Euclidean representation

A *Cartesian space*, named after René Descartes, French mathematician and philosopher (1596-1650), is a space with a Cartesian system of coordinates. It is also called a *Euclidean space* because the distances among points are measured by equation 6 in that space. The multidimensional ordination spaces of PCA, CA, PCoA, NMDS, etc., are Cartesian or Euclidean spaces; hence the distances among points embedded in these spaces are measured by the Euclidean distance formula. A few dimensions that represent a good deal of the variance of the data will be chosen from these multidimensional spaces to create a reduced-space ordination.

A distance function is said to have the Euclidean property, or (in short) to be Euclidean, if it always produces distance matrices that are fully embeddable in a Euclidean space. The test, available in the R-language package ADE4 (function ‘is.euclid’), is that a principal coordinate analysis (PCoA) of a Euclidean distance matrix produces no negative eigenvalues. This is not always the case in ordination. Some distance functions are not Euclidean, meaning that the distances in the matrix cannot be fully represented in a Euclidean ordination space. A principal coordinate analysis of the distance matrices produced by these coefficients may generate negative eigenvalues; these eigenvalues indicate the non-Euclidean nature of the distance matrix (Gower 1982). They measure the amount of variance that needs to be added to the distance matrix to obtain a full Euclidean representation. To be a metric is a necessary but not a sufficient condition for a distance coefficient to be Euclidean. Many of the commonly-used similarity coefficients are not Euclidean when transformed into distances using equation 13. The transformation described by equation 14 often solves the problem, however. For instance, the similarity coefficients of Gower, simple matching, Jaccard, Sørensen, Ochiai and Steinhaus, described above, all become Euclidean when transformed into distances using equation 14 (Gower and Legendre 1996; Legendre and Legendre 1998, Table 7.2).

If the analysis is carried out to produce a PCoA ordination in a few (usually 2 or 3) dimensions, negative eigenvalues do not matter as long as their absolute values are not large when compared to the positive eigenvalues used for the reduced-space ordination. If the analysis requires that all coordinates be kept, as it will be the case when testing multivariate hypotheses using the db-RDA method (see the subsection below on Linear RDA), negative eigenvalues should either be avoided or corrected for. They can be avoided by selecting a distance coefficient that is known to be Euclidean. When a non-Euclidean coefficient is used (for example, the Steinhaus/Odum/Bray-Curtis coefficient of equation 21), there are ways of correcting for negative eigenvalues in PCoA to obtain a fully Euclidean solution; see Legendre and Legendre (1998: 432-438) for details. These corrections are available in some PCoA computer programs, including function 'pcoa' of the APE R-language package.

Metric or non-metric ordination?

Metric ordinations are obtained by the methods of principal component analysis (PCA), correspondence analysis (CA), and principal coordinate analysis (PCoA). These methods all proceed by eigenvalue decomposition. The eigenvalues measure the amount of variation of the observations along the ordination axes. The distances in the full-dimensional ordination space are projected onto the space of reduced dimensionality (usually 2-dimensions) chosen for ordination. Non-metric ordinations are obtained by non-metric multidimensional scaling (NMDS) which is not an eigenvalue method. This method only preserves the rank-order of the original distances in the reduced ordination space.

PCA is the method of choice to preserve Euclidean distances among objects, and CA when the chi-square distance is to be preserved. For other forms of distance, users have to choose between PCoA (also called metric scaling) and NMDS. PCoA is the preferred method (1) when one wishes to preserve the original distances in full-dimensional space, (2) when many (or all) ordination axes are sought, or (3) when the data-set is fairly large. NMDS may be preferred when the user wants to represent as much as possible of the distance relationships among objects in a few dimensions, at the cost of preserving only the rank-order of the distances and not the distances themselves.

The size of the data-sets is also of importance. PCA and CA can easily be computed on very large data-sets (tens or hundreds of thousand objects) as long as the number of variables is small (up to a few hundred), because the eigenvalue decomposition is done on the covariance matrix, which is of size *p*, the number of variables in the data-set.

For tables containing assemblage composition data, three paths can be followed: (1) one can transform the data using one of the transformations described by equations 5, 7, 8, or 9, and produce the ordination by PCA (tb-PCA approach), or (2) compute a distance matrix using equations 15 to 24, followed by PCoA or NMDS. For large data-sets of intermediate sizes (a few hundred objects), PCoA will produce the ordination solution faster than NMDS. For very large data-sets, PCA should be used. (3) For data-sets of any size, one can produce the ordination using CA if the chi-square distance is appropriate.

An alternative and biologically useful approach to deciding between ordinations based on PCA (Euclidean distance) of untransformed data and CA (chi-square distance) of multivariate species assemblage data is that emphasised by ter Braak and Prentice (1988) and ter Braak (1987c), namely the underlying species response model that is assumed when fitting either PCA or CA and extracting synthetic latent variables that are then used as the major ordination axes. PCA assumes an underlying linear response model, whereas CA assumes an underlying unimodal response model between the variables and the unknown but to be determined latent variables or ordination axes. The question is thus how to know whether a linear-based or a unimodal-based ordination is appropriate for a given data-set. The detrended relative of CA, detrended correspondence analysis (DCA: Hill and Gauch 1980; ter Braak 1987c), is a heuristic modification of CA designed to minimise two of the disadvantages of CA, namely the so-called arch-effect and the so-called edge-effect (ter Braak and Prentice 1980). As a result of the non-linear rescaling of the axes that removes the edge-effect, the object scores are scaled and standardised in a particular way. The lengths of the resulting ordination axes are given by the range of object scores and are expressed in “standard deviation units” (SD) or units of compositional turnover. The tolerance or amplitude of the species’ curves along the rescaled DCA axes are close to 1; each curve will therefore rise and fall over about 4 SD (ter Braak and Prentice 1985). Objects that differ by 4 SD can be expected to have no species in common. A preliminary DCA of an assemblage data-set, with detrending by segments and non-linear rescaling, provides an estimate of the underlying gradient length. If the gradient length is less than about 2.5 SD, the assemblage variation is within a relatively narrow range, and the linear approach of PCA is appropriate. If the gradient length is 3 or more SD, the assemblage variation is over a larger range, and the unimodal-based approach of CA is appropriate (ter Braak and Prentice 1985). Transformation-based PCA (tb-PCA) is also appropriate in that case.

How many axes are required?

In most instances, ordination analysis is carried out to obtain an ordination in 2, sometimes 3, dimensions. The ordination is then used to illustrate the variability of the data along the ordination axes and attribute it to the variables that are most highly correlated with those axes. Simple interpretation of the variability in the ordination diagram can be obtained by projecting interpretative variables in the ordination plane, or by representing other properties of the data (for instance, the groups produced by cluster or partitioning analysis), or some other grouping of the objects known a priori (for example, the type of lake, or the nature of the sediment) (see Lepš and Šmilauer 2003).

There are instances where ordination analysis is carried out as a pre-treatment, or transformation, of the original data, before carrying out some further analysis. For example, one may wish to preserve the Steinhaus/Odum/Bray-Curtis distance in a canonical redundancy analysis (RDA) or *K*-means partitioning (see Legendre and Birks: Chapter 6 this volume). To achieve that, one may compute the distance matrix using equation 22 (or its square root) and carry out a PCoA of that matrix. One then keeps all eigenvectors from this analysis (after perhaps a correction for negative eigenvalues) and uses that matrix of eigenvectors as input to redundancy analysis (RDA) or *K*-means partitioning. This is an example of distance-based RDA (db-RDA) described in more detail in the subsection on Linear RDA.

Tests of significance for individual eigenvalues are available for PCA; see the review papers of Burt (1952) and Jackson (1993). They are not often useful because, in most instances, ecologists do not have a strong null hypothesis to test; they rather use PCA for an exploratory representation of their data. Also, the parametric tests of significance assume normality of all descriptors, which is certainly a drawback for palaeolimnological data. Users most often rely on criteria that help them determine how many axes represent "important" variation with respect to the original data table. The two best criteria at the moment are the simple broken-stick model proposed by Frontier (1976) as well as the bootstrapped eigenvalue method proposed by Jackson (1993).

Simple ordination methods: PCA, CA, PCoA, NMDS

The simple ordination methods mostly used by palaeoecologists and palaeolimnologists (Table 1) are the following.

1) Principal component analysis (PCA) is the oldest (Hotelling 1933) and best-known of all ordination methods. Consider a group of data points in multidimensional space, placed at Euclidean distances (equation 6) of one another. Imagine a lamp behind the cloud of points, and the shadows of the points projected onto a white wall. The geometric problem consists of rotating the points in such a way that the shadows have as much variance as possible on the wall. The mathematics of eigenvalues and eigenvectors, which is part of matrix algebra, is the way to find the rotation that maximises the variance of the projection in any number of dimensions. The variables are first transformed if required (Table 2), then centred by column, forming matrix $\mathbf{Y}$. One computes the dispersion (or variance-covariance) matrix $\mathbf{S}$ among the variables, followed by the eigenvalues (λ_j) and eigenvectors of $\mathbf{S}$. The eigenvectors are assembled in matrix $\mathbf{U}$. The principal components, which provide the coordinates of the points on the successive ordination axes, are the columns of matrix $\mathbf{F} = \mathbf{YU}$. The eigenvalues measure the variance of the points along the ordination axes (the columns of matrix $\mathbf{F}$). The first axis has the highest eigenvalue λ_1, hence the largest variance; and so on for the following axes, with the constraint that all axes are orthogonal and uncorrelated to one another.

A scatter diagram with respect to the first 2 ordination axes, using the coordinates in the first 2 columns of matrix $\mathbf{F}$, accounts for an amount of variance equal to $\lambda_1 + \lambda_2$. The distances among points in 2 dimensions are projections of their original, full-dimensional Euclidean distances. The contributions of the variables to the ordination diagram can be assessed by drawing them using the loadings found in matrix $\mathbf{U}$. For 2 dimensions again, the first 2 columns of matrix $\mathbf{U}$ provide the coordinates of the end-points of vectors representing the successive variables. A graph presenting the variables (as arrows) on top of the dispersion of the points, as described above, is called a *distance biplot*. Another type of biplot, called a *correlation biplot*, can also be produced by many PCA programs; the correlations among variables are represented by the angles of projection of the variables, in 2 dimensions, after rescaling the eigenvectors to the square root of their respective eigenvalues (ter Braak 1994; Lepš and Šmilauer 2003). Supplementary or passive objects and variables can be projected onto a PCA ordination diagram. This option is available in most of the programs offering a PCA procedure listed in Table 3. The mathematics behind such projections is described in Legendre and Legendre (1998, Section 9.1.9) and ter Braak and Šmilauer (2002).

The approach of fitting fossil objects as supplementary objects onto a PCA ordination has been used by palaeoecologists (e.g., Lamb 1984) as an aid in detecting similarities between modern and fossil assemblages. It is important, however, to calculate the residual distances when adding additional supplementary objects into any low-dimensional ordination, as new objects may appear to be positioned close to other objects on the first few axes and yet be located some distance from the other objects when further dimensions are considered (Birks and Gordon 1985). Gower (1968) discusses the calculation and interpretation of the residual distances from the true position of the added points to the fitted plane giving the best two-dimensional representation of the objects.

Alternatively, one may perform a PCA of fossil assemblage data and add modern objects into the ordination (e.g., Ritchie 1977), or perform a PCA of fossil and modern assemblage data combined (MacDonald and Ritchie 1986). Prentice (1980) and Birks and Gordon (1985) discuss the advantages and disadvantages of fitting objects, modern or fossil, into low-dimensional PCA representations.

The most common application of PCA in palaeolimnology is to produce biplot diagrams of the objects (sites, lakes, core subunits, etc.) with respect to physical or chemical variables (e.g., Jones et al. 1993) or assemblage composition data (after appropriate transformation: Table 1) (e.g., Birks and Peglar 1979). Another useful representation of PCA results of core assemblages is to plot the object scores on the first few principal components in stratigraphical order for each axis (e.g., Birks and Berglund 1979; Birks 1987; Lotter and Birks 2003; Birks: Chapter 9 this volume), thereby providing a summarisation of the major patterns of variation in the stratigraphical data in 2 or 3 axes. PCA can also be used to detect outliers in data, which may correspond to legitimate outliers, or to erroneous data. PCA may be used to identify groups of variables that are highly correlated and, thus, form bundles of arrows in the ordination diagram; look, in particular, for variables that are highly but negatively correlated: their arrows are opposite in the diagram (e.g., Gordon 1982; MacDonald and Ritchie 1986). Another application is to simplify data-sets containing many highly collinear variables; the PCA axes that account for, say, 95% of the total variance form a simplified set of variables and allow discarding of the remaining 5%, which can be regarded as noise (Gauch 1982; Lotter et al. 1992).

2) Correspondence analysis (CA) is a form of PCA that preserves the chi-square distance (equation 19) among the objects or variables. CA is appropriate for frequency data, and in particular for species presence-absence or abundance data, subject to the caveat that the chi-square distance gives high weights to rare species. There are several ways of presenting CA (Hill 1974). We will look at it here as the eigenanalysis of a table of components of chi-square. The assemblage composition data matrix $\mathbf{Y}$ is transformed into a matrix of components of chi-square $\mathbf{Q} = [q_{ij}]$ where

$$q_{ij} = \left[\frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}\right] \Big/ \sqrt{y_{++}} \qquad (25)$$

The part inside the square parentheses is easily recognised as the component of the chi-square statistic computed in each cell of a frequency (or contingency) table; they are obtained from the observed (O_{ij}) and the expected values (E_{ij}) of cell *ij* of the table. These components can be added to produce the Pearson chi-square statistic used to test the hypothesis of absence of relationship between the rows and columns of a contingency table. Here, the components of chi-square are divided by a constant, the square root of the sum of values in the whole table (y_{++}), which turns them into the values [q_{ij}] of the transformed data table **Q**. From this point, one can compute a cross-product matrix (the covariance matrix computed in PCA is also a cross-product matrix, but it is computed here without further centring since centring is part of equation 25), and from it the eigenvalues and eigenvectors. An alternative approach is to carry out singular value decomposition of the matrix **Q**, as explained in Legendre and Legendre (1998, Section 9.4). The eigenvalues measure the amount of inertia accounted for by each ordination axis. Matrices are obtained that contain the positions of the objects (rows) and species (columns) along the successive axes of the ordination space. Two types of scaling can be used for biplots: one can (1) preserve the chi-square distances among objects (rows), the objects being at the centroids of the species (columns); or (2) preserve the chi-square distance among the variables (columns), the variables being at the centroids of the objects (rows) (ter Braak and Verdonschot 1995). The most common application of CA in palaeolimnology is to produce biplot diagrams of species and objects or other sampling units (e.g., Jones and Birks 2004). As in PCA, supplementary objects and variables can be projected onto a CA ordination diagram (e.g., Jones and Birks 2004). This option is available, for instance, in the program CANOCO. In R, functions to that effect are also available in VEGAN and ADE4. *** Should we list the functions? The list is rather long! vegan: ***predict.rda()*** and ***predict.cca()*** for adding new points to PCA, RDA, CA and CCA, and ***envfit()*** for adding supplementary variables to all of the above (envfit does weighted fitting in CCA so that it is consistent with the original). ade4: ***suprow()*** to add supplementary objects and ***supcol()*** to add supplementary variables to PCA and CA plots. ***

Usually, ecologists who see the organisms they are sampling consider rare species as potential indicators of rare environmental conditions, whereas those who have to sample blindly or use traps are more wary of the interpretation of rare species. In animal ecology, a single presence of a species at a site may be due to a species that does not belong to the site but was travelling between two other favourable sites. In palynology, likewise, pollen may be brought by far transport from distant sites. In aquatic ecology, rare species may appear in spurious ways in sampling units from sites where they are found at low abundance. Because of their influence on the chi-square distance (equation 19), one should pay special attention to rare species in CA. One must understand that rare species affect the ordination of objects very little, but these species will be represented by points located far from the origin. Users of CA who are worried about the interpretation of rare species often decide to remove, not the species that have low abundance, but those that occur in the data-set very rarely. One may try removing first the species that occur only once in the data-set, then those that occur once or twice, and so on, recomputing the analysis every time. One can remove the rarest species up to the point where the first few eigenvalues, expressed as percentages of the inertia *in* *the original data-set*, are little affected by the removal. This approach has been suggested by Daniel Borcard, Université de Montréal.

Palaeolimnologists often use the detrended relative of CA, detrended correspondence analysis (DCA) as a preliminary tool in establishing the extent of compositional turnover in modern calibration data-sets as a guide as to whether to use calibration procedures that assume linear or unimodal responses of species to environmental gradients (Birks 1995). Detrending by segments is an arbitrary method for which no theoretical justification has been offered, while the assumptions behind the nonlinear rescaling procedure have not been fully substantiated (Wartenberg et al. 1987, but see ter Braak 1985). Jackson and Somers (1991) showed that DCA ordinations of sites greatly varied with the number of segments one arbitrarily decides to use, so that the ecological interpretation of the results may vary widely. In simulation studies conducted on artificial data representing unimodal species responses to environmental gradients in one or two dimensions, DCA did not perform particularly well in recovering complex gradients (Kenkel and Orlóci 1986, Minchin 1987). For these reasons, detrended correspondence analysis (DCA) should generally be avoided for the production of ordination plots.

3) In principal coordinate analysis (PCoA), the objective is to obtain an ordination, in any number of dimensions, representing as much as possible of the variation of the data while preserving the distance that has explicitly been computed. The algebra used to find a solution to the geometric problem proceeds directly from a pre-computed square, symmetric distance matrix **D**. The first step is to transform the distances d_{hi} of **D** into values $a_{hi} = -0.5d^2_{hi}$, then to centre the resulting matrix **A** to produce a third matrix $\mathbf{\Delta} = [\delta_{hi}]$ using the equation:

$$\delta_{hi} = a_{hi} - \bar{a}_h - \bar{a}_i + \bar{a} \tag{26}$$

where $\bar{a}_h$ and $\bar{a}_i$ are the means of row h and column i corresponding to element a_{hi}, whereas $\bar{a}$ is the mean of all a_{hi} values in the matrix. Eigenvalue decomposition is applied to matrix $\mathbf{\Delta}$, producing eigenvalues and eigenvectors. When the eigenvectors are normalised to the square root of their respective eigenvalues, they directly provide the coordinates of the objects on the given ordination axis. The eigenvalues give the variance (not divided by degrees of freedom) of the objects along that axis. If some eigenvalues are negative and all ordination axes are needed for subsequent analyses, corrections can be applied to the distance matrix; this was mentioned in the section on Euclidean or Cartesian space, Euclidean representation.

A simple example may help explain PCoA. From an object-by-variable data matrix **Y**, compute matrix **D** of Euclidean distances among the objects. Run PCA using matrix **Y** and PCoA using matrix **D**. The eigenvalues of the PCoA of matrix **D** are proportional to the PCA eigenvalues computed for matrix **Y** (they differ by the factor (n – 1)), while the eigenvectors of the PCoA of **D** are identical to matrix **F** of the PCA of **Y**. Normally, one would not compute PCoA on a matrix of Euclidean distances since PCA is a faster method to obtain an ordination of the objects in **Y** that preserves the Euclidean distance among the objects. This was presented here simply as a way of understanding the relationship between PCA and PCoA in the Euclidean distance case. The real interest of PCoA is to obtain an ordination of the objects from some other form of distance matrix more appropriate to the data at hand—for example, a Steinhaus/Odum/Bray-Curtis distance matrix in the case of assemblage composition data.

Surprisingly, PCoA has rarely been used in palaeoecology (e.g., Birks 1977) in contrast to the extensive use of PCA, CA, and DCA.

4) Non-metric ordinations are obtained by non-metric multidimensional scaling (NMDS); several variants of this method have been proposed (Prentice 1977, 1980). The distances in the low-dimensional space are not rigid projections of the original distances in full-dimensional space. In NMDS, the user sets the dimensionality of the space in which the ordination is to be produced; the solution sought is usually two-dimensional. The program proceeds by successive iterations, trying to preserve in the ordination the rank-order of the original distances. Different functions, called Stress (formula 1 or 2), Sstress, or Strain, may be used to measure the goodness-of-fit of the solution in reduced space. Non-metric ordinations are rarely used in palaeoecology. Early applications include Birks (1973), Gordon and Birks (1974), and Prentice (1978), whereas more recent applications include Brodersen et al. (1998, 2001) and Simpson et al. (2005). Overall, there seem to be few theoretical advantages in using NMDS in palaeoecology (Prentice 1980).

Introduction to canonical ordination

The methods of canonical ordination are generalisations of simple ordination methods; the ordination is forced or constrained to represent the part of the variation in a table of response variables (e.g., species abundances) that is maximally related to a set of explanatory variables (e.g., environmental variables). Canonical redundancy analysis (RDA) is the constrained form of PCA whereas canonical correspondence analysis (CCA) is the constrained form of CA. Canonical ordination is a hybrid between regression and ordination, as will be described below. The classical forms of RDA and CCA use multiple linear regression between the variables in the two data tables. Canonical ordination methods have also been described that look for polynomial relationships between the dependent (response) and explanatory (predictor) variables. Tests of statistical significance of the relationship between the species and environmental data can be performed in canonical ordination, just as in multiple regression.

Canonical ordination methods are widely used in palaeolimnological studies. The Birks et al. (1998) bibliography on the use of canonical analysis in ecology for the period 1986-1996 contained 804 titles, 96 of which are in the fields of palaeobotany, palaeoecology, and palaeolimnology. Applications of these methods in palaeoecology (Table 1) try to establish links between species assemblages and environmental factors, or use canonical analysis as a first step in calibration studies to guide the selection of significant factors that may be estimated by taxonomic assemblages (Birks 1995) (see Juggins and Birks: Chapter 12 this volume). Palaeolimnologists also try to estimate how much of the assemblage variation can be attributed to different groups of environmental factors, such as sediment types, geology, climatic factors, geography, topography, land-use, etc. (e.g., Lotter et al. 1997; Simpson and Hall: Chapter 16 this volume)).

Canonical ordination methods

The types of canonical ordination methods that palaeoecologists are mostly interested in are redundancy analysis (RDA) and canonical correspondence analysis (CCA). They are asymmetric forms of analysis, combining regression and ordination. These analyses focus on a clearly identified table of response variables (containing, very often, assemblage composition data), which is related to a table of explanatory variables (e.g., environmental variables). Other forms of canonical analysis are available in the major statistical packages: canonical correlation analysis (CCorA) and canonical variate analysis (CVA), also called multiple discriminant analysis (see ter Braak 1987c). These forms will not be discussed in this chapter because they do not treat matrix $\mathbf{Y}$ as a response data table; they are briefly outlined in Chapter 2 (this volume). Other more general approaches to the linking of two or more ecological data tables are co-inertia analysis (Dolédec and Chessel 1994, Dray et al. 2003) and multiple factor analysis (Escofier and Pagès 1994); they allow the analysis of a wide range of different data tables (Dray et al. 2003), with no constraints on the number of species and environmental variables in relation to the number of objects or on the role of the different tables as response and predictor variables. All these methods of canonical analysis are described and illustrated in Chapter 6 of Borcard et al. (2011).

In the asymmetric forms of canonical analysis, after regressing the $\mathbf{Y}$ variables on $\mathbf{X}$, an ordination is computed on the regression fitted values. The preliminary questions that have to be resolved before ordination (Table 1) will also have to be answered about the data in $\mathbf{Y}$ prior to canonical ordination: the choice of transformations for the physical or species data, and of an appropriate distance measure among objects. The table of explanatory variables, called $\mathbf{X}$, contains the independent (or constraining) variables used in the regression part of the analysis. The decisions normally made prior to or during regression will have to be considered prior to canonical analysis: transformation of the regressors; coding of multi-state qualitative variables into dummy (binary or orthogonal) variables; coding the factors of experiments into (orthogonal) dummy variables; and choice of a linear or polynomial regression model. We do not have to worry about (multi)normality since the tests of significance in canonical analysis are carried out by Monte Carlo permutation tests (see Legendre and Legendre 1998, 2012; Lepš and Šmilauer 2003; Birks: Chapter 2 this volume; Lotter and Anderson: Chapter 15 this volume).

Linear RDA

Canonical redundancy analysis (RDA) combines two steps: linear regression and PCA. The analysis is schematically described in Figure 2. (1) Each variable (column) of $\mathbf{Y}$ is regressed on $\mathbf{X}$, which contains the explanatory variables. The fitted values of the multiple regressions are assembled in matrix $\hat{\mathbf{Y}}$, whereas the residuals are placed in the columns of matrix $\mathbf{Y}_{\mathbf{res}}$. $\hat{\mathbf{Y}}$ thus contains that part of $\mathbf{Y}$ that is explained by linear models of $\mathbf{X}$, whereas $\mathbf{Y}_{\mathbf{res}}$ contains that part of $\mathbf{Y}$ that is linearly independent of (or orthogonal to) $\mathbf{X}$. At this point, the matrices $\hat{\mathbf{Y}}$ and $\mathbf{Y}_{\mathbf{res}}$ have the same number of columns as $\mathbf{Y}$. (2) The matrix of fitted values $\hat{\mathbf{Y}}$ usually contains (much) less information, measured by its total variance, than $\mathbf{Y}$. A PCA of $\hat{\mathbf{Y}}$ is computed to reduce its dimensionality, producing eigenvalues (that are now called canonical eigenvalues), a matrix of eigenvectors $\mathbf{U}$ (now called canonical eigenvectors, which will be used as the matrix of response variable scores for the biplot), and a matrix **Z** of principal components, obtained in the same way as matrix **F** of the principal components in PCA, which contains the sampling unit scores for the ordination biplot; for details, refer to the description of PCA in the previous section on Simple ordination methods. In some applications, ecologists prefer to use, for biplots, the sampling unit scores obtained by the operation $\mathbf{F} = \mathbf{YU}$ (upper-right in Figure 2). These scores are not the direct result of the PCA of the fitted values $\hat{\mathbf{Y}}$; they are based on the original data **Y**, which contain the fitted values plus the residuals (noise). These sampling unit scores (column vectors of matrix **F**) are not orthogonal to each another, since they differ from the vectors of matrix **Z**, which are orthogonal as in any PCA. (3) In some applications, the effect of the explanatory variables on **Y** is already well documented and understood; for instance, the effect of water depth on aquatic macroinvertebrates. RDA can be used to go beyond what is already known, by examining the residuals of the regression, found in matrix $\mathbf{Y}_{\mathbf{res.}}$ In those cases, one is interested in obtaining an ordination of the matrix of residual variation: a PCA is performed on matrix $\mathbf{Y}_{\mathbf{res}}$, as shown in the lower part of Figure 2.

Scalings in RDA biplots follow the same rules as in PCA: one may be primarily interested in an ordination preserving the Euclidean distances among sampling unit fitted values (distance biplot), or in illustrating the correlations among the columns of $\hat{\mathbf{Y}}$ (correlation biplot) (ter Braak 1994). The explanatory environmental variables can also be represented in the ordination diagrams, which become triplots, by using their correlations with the canonical ordination axes. The correlation coefficients must be slightly modified to account for the stretching of the canonical ordination axes; the biplot scores of environmental variables are obtained by multiplying the correlation coefficients by $\sqrt{}(\lambda_k$/total variance in **Y**). States of binary or multistate qualitative variables can be usefully represented in triplots by the centroids (mean coordinates) of the sampling units that possess the given state (ter Braak 1994).

The number of canonical axes is limited by either the number of variables in **Y** or the number of variables in **X**. Example 1: if **Y** contains a single variable, regressing it on **X** produces a single vector of fitted values and, hence, a single canonical axis. Example 2: if **X** contains a single column, regressing **Y** (which contains p columns) on **X** will produce a matrix $\hat{\mathbf{Y}}$ with p columns, but since they are the result of regression on the same explanatory variable, matrix $\hat{\mathbf{Y}}$ is actually one-dimensional. So, the PCA will come up with a single non-zero eigenvalue that will contain all the variance of **Y** explained by **X**. The analysis of a matrix $\mathbf{Y}(n \text{ x } p)$ by a matrix $\mathbf{X}(n \text{ x } m)$ produces at most $(n–1)$, p, or m canonical axes, whichever is the minimum.

Like PCA, RDA can be tricked into preserving some distance that is appropriate to assemblage composition data, instead of the Euclidean distance (Figure 3). Figure 3b shows that assemblage composition data can be transformed using equations 5 or 7-9 (transformation-based RDA, or tb-RDA, approach). RDA computed on data transformed by these equations will actually preserve the chord, chi-square, profile, or Hellinger distance among sites. One can also (Figure 3c) directly compute one of the distance functions appropriate for assemblage composition data (equations 15-24), carry out a principal coordinate analysis of the distance matrix, and use all the PCoA eigenvectors as input to RDA. This is the distance-based RDA approach (db-RDA) advocated by Legendre and Anderson (1999).

Partial RDA offers a way of controlling for the effect of a third data-set, called the matrix of covariables **W**. Computationally, the analysis first calculates the residuals $\mathbf{X}_{\mathbf{res}}$ of the explanatory variables **X** on the covariables **W**;

then an RDA of $\mathbf{Y}$ on $\mathbf{X}_{\mathbf{res}}$ is computed; see details in Legendre and Legendre (2012). This is quite different from a PCA of $\mathbf{Y}_{\mathbf{res}}$ mentioned at the end of the introductory paragraph of the present section. Partial RDA is a generalisation of partial linear regression to multivariate data, for example, species assemblages. It is used in many different situations, including the following: (1) controlling for the effect of $\mathbf{W}$ (e.g., geographic positions) in tests of the relationship between $\mathbf{Y}$ (e.g., modern biological assemblages) and $\mathbf{X}$ (e.g., modern environmental data) (Peres-Neto and Legendre 2010); (2) determining the partial, singular effect of an explanatory variable of interest (e.g., environmental), and testing its significance, while controlling for the effect of all the other explanatory variables in the study; (3) partial RDA is used to test the significance of single factors and interaction terms in two-way or multi-way experimental designs where species assemblages are the response variable (see Testing hypotheses in (multi-)factorial experiments below); (4) partial RDA is also used to test the significance of individual fractions in variation partitioning (see Spatial or temporal analysis through variation partitioning below).

In terms of algorithms, RDA and CCA can be obtained either by global regression and PCA, as described here, or by the iterative algorithm described by ter Braak (1987c) and used in the CANOCO program. In large analyses, the global algorithm produces more precise results when many canonical ordination axes are to be extracted and used in further analyses; the iterative algorithm is computationally faster when one is only interested in obtaining the first few (4-8) canonical axes.

Linear CCA

Canonical correspondence analysis (CCA) only differs from RDA in two aspects. First, it is the matrix $\mathbf{Q}$ of CA (see Simple ordinations methods above) which is used as the response data matrix, instead of the data matrix $\mathbf{Y}$. This ensures that the chi-square distance is preserved by CCA among the rows of the response data table and the assumption of unimodal species responses is made as in CA. Second, the regression step is carried out using weights p_{i+}; p_{i+} is the sum of frequencies in row i (y_{i+}) divided by the grand total (y_{++}) of all frequencies in the table. Using these weights is tantamount to repeating each row of the response and explanatory data tables y_{i+} times before computing the regressions. Scalings for biplots or triplots are the same as in CA (see ter Braak and Verdonschot 1995). Just as one can compute a partial RDA, it is possible to perform a partial CCA (ter Braak and Prentice 1988). Odgaard (1994) provides an illustrative application of partial CA in palaeoecology and Bradshaw et al. (2005) provide a detailed application of partial CCA in palaeolimnology.

Fossil assemblages can be positioned as supplementary or passive objects in a CCA or RDA of modern biological assemblages, in relation to modern environmental variables, to provide a projection of fossil samples (from an unknown past environment) into modern 'environment–species–object' space (e.g., Birks et al. 1990a; Allott et al. 1992; Juggins and Birks: Chapter 12 this volume; Simpson and Hall: Chapter 16 this volume).

There have been many applications of RDA and CCA and their partial forms in palaeoecology and palaeolimnology in either a descriptive mode to display modern species–object–environment relationships (e.g., Birks et al. 1990a) or in an analytical, hypothesis-testing mode. Illustrative examples of the latter approach include Lotter and Birks (1992), Renberg et al. (1993), Korsman et al. (1994), Anderson et al. (1995), Korsman and

Segerström (1998), Odgaard and Rasmussen (2000), and Bradshaw et al. (2005) (see Lotter and Anderson: Chapter 15 this volume).

Other forms of asymmetric canonical analyses

There is no special reason why nature should linearly relate changes in community composition to changes in environmental variables. While they know that the assumption of linearity is often unrealistic, users of RDA and CCA sometimes used the linear forms of these methods simply because more appropriate models were not available. Makarenkov and Legendre (2002) proposed a nonlinear form of RDA and CCA, based on polynomial regression, to do away with the assumption of linearity in modelling the relationships between tables of response and explanatory variables. Their algorithm includes a stepwise procedure for selection of the best combination of linear and quadratic terms of the explanatory variables.

Palaeolimnologists may want to relate two types of assemblages, for example predators and preys. For a top-down model, the predators would form the data-set explaining the variation of the prey, and the opposite for a bottom-up model. To relate two communities, ter Braak and Schaffers (2004) proposed a model of co-correspondence analysis. An alternative method is to transform the two community data tables using one of the transformations described in the section below on Transformation of community composition data, as proposed by Pinel-Alloul et al. (1995), and analyse the two tables using RDA. As noted by ter Braak and Schaffers (2004), one should not use forward selection of the species in the explanatory table during this type of analysis. The R-package COCORRESP (Simpson 2009) implements co-correspondence analysis.

Spatial or temporal analysis through variation partitioning

Variation partitioning is an approach to the analysis of a response variable or data table, using two or more explanatory variables or data tables. For simple response variables, the analysis is carried out using partial linear regression; see Legendre and Legendre (1998, Section 10.3.5). Partial canonical analysis, which is available in CANOCO and VEGAN, allows ecologists to partition the variation of a response data table among two explanatory tables, using RDA or CCA.

In the original proposal (Borcard et al. 1992), the proportion of variation of a response variable or data table accounted for by a table of explanatory variables was estimated using the ordinary coefficient of determination (R^2). It has long been known that R^2 is a biased estimator of the proportion of explained variation. Ohtani (2000, for regression) and Peres-Neto et al (2006, for canonical analysis) have shown that the adjusted coefficient of determination (R_a^2 , Ezekiel 1930),

$$R_a^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-m-1}\right) \qquad (27)$$

is unbiased, where n is the number of observations and m is the number of explanatory variables. Peres-Neto et al. (2006) have also shown how to compute the fractions of variation described in the next paragraph using R_a^2 . The R-language function 'varpart' available in the VEGAN package allows users to partition the variation of a response data table **Y** among 2, 3, or 4 tables of explanatory variables $\mathbf{X}_1$ to $\mathbf{X}_4$.

The variation-partitioning approach was first advocated by Borcard et al. (1992) in the context of spatial analysis in which a species composition response table **Y** is partitioned between a matrix of environmental variables and one describing the spatial relationships among the sampling sites. The variation in **Y** is partitioned into four fractions, three of which can be interpreted separately or in combinations (Figure 4): [a] is the non-spatially-structured component of the variation of **Y** explained by the environmental variables, [b] is the spatially-structured component explained by the environmental variables, [c] is the amount of spatially-structured variation of **Y** not explained by the environmental variables used in the analysis, and [d] is the unexplained (residual) variation.

In Borcard et al. (1992) and Borcard and Legendre (1994), as well as in the many applications published since 1992, the spatial relationships were represented in the analysis by a polynomial of the geographical coordinates of the sampling sites. A new form of spatial partitioning, based on Principal Coordinates of Neighbour Matrices (PCNM), has been proposed by Borcard and Legendre (2002). In PCNM analysis, the polynomial function of the geographic coordinates of the sites of Borcard et al. (1992) is replaced by a set of spatial eigenfunctions, the PCNMs, corresponding to a spectral decomposition of the spatial relationships among the sites. PCNM analysis allows the modelling of spatial or temporal relationships at all spatial scales that can be perceived by the sampling design. Borcard et al. (2004) and Legendre and Borcard (2006) present several applications to the analysis of multivariate spatial patterns. Telford and Birks (2005) have also applied PCNM analysis to explore the spatial structures within core-tops of foraminiferal assemblages in the Atlantic.

Dray et al. (2006) examined the link between PCNM analysis and spatial autocorrelation structure functions, and generalised the method to different types of spatial weightings. The generalised eigenfunctions are called Moran's Eigenvector Maps (MEM). Griffith and Peres-Neto (2006) unified Dray's MEM spatial eigenfunctions with Griffith's (2000) spatial eigenfunctions. Blanchet et al. (2008) developed Asymmetric Eigenvector Maps (AEM) to model species spatial distributions generated by hypothesised directional physical processes such as migrations in river networks and currents in water bodies.

Several R-language functions are available to compute PCNM and MEM spatial eigenfunctions: 'pcnm' in VEGAN, 'pcnm' in the SPACEMAKER package, 'PCNM' and 'quickPCNM' in the PCNM package; the last two libraries are available at http://r-forge.r-project.org/R/?group_id=195. Several applications of spatial eigenfunction analysis to ecological data in R are presented by Borcard et al. (2010). A stand-alone program called 'SpaceMaker2' is also available at http://www.bio.umontreal.ca/legendre/indexEn.html to compute PCNM eigenfunctions.

Modelling temporal structure in sediment cores [and environmental structure in modern assemblages]

Variation partitioning of stratigraphical palaeolimnological data has been used in various studies to partition variation in a biostratigraphical sequence (e.g., diatoms) into components explained by the occurrence of volcanic ash or other potential perturbations, by climatic changes, and by natural temporal shifts (e.g., Lotter and Birks, 1993, 1997, 2003; Lotter et al. 1995; Barker et al. 2000; Eastwood et al. 2002). It has also been used to partition variation in modern biological assemblages (e.g., diatoms) in relation to a range of explanatory variables such as lake-water chemistry, climate, geography, etc. (e.g., Jones and Juggins 1995; Pienitz et al. 1995; Gasse et al 1995; Lotter et al. 1997, 1998; Potopova and Charles 2002; Simpson and Hall: Chapter 16 this volume) and in fossil assemblages in relation to spatial and temporal variables (e.g., Ammann et al. 1993). Variation partitioning is being increasingly applied as a hypothesis-testing approach in palaeolimnology, to quantify the proportion of total variation in assemblage composition over time explicable by various environmental variables. For example, Hall et al. (1999) used high-resolution diatom core data and 100 years of historical data to quantify the effects of climate, agriculture, and urbanisation on diatom assemblages in lakes in the northern Great Plains. They showed that human impact was the major determinant of biotic change. Quinlan et al. (2002) obtained similar results for the same area using fossil chironomid assemblages. The use of variation partitioning requires careful project design to exploit 'natural experiments' (e.g., factorial designs) and to test critical hypotheses. Other detailed palaeolimnological applications of variation partitioning to test specific hypotheses include Vinebrooke et al. (1998) and Leavitt et al (1999). Birks (1998) reviewed the use of variation partitioning as a means of testing hypotheses in palaeolimnology (see also Lotter and Anderson: Chapter 15 this volume).

We now present an example to illustrate the use of canonical ordination as a form of spatial or time-series analysis for multivariate ecological response data. The Round Loch of Glenhead (RLGH) fossil data consist of the counts of 139 Holocene diatom taxa observed in 101 levels of a sediment core from a small lake in Galloway, south-western Scotland (Jones et al. 1989; see Birks and Jones: Chapter 3 of this volume). The data series covers the past 10,000 years. Level no. 1 is the top (most recent), no. 101 is the bottom of the core (oldest). The diatom counts were expressed as proportions relative to the total number of cells in each section of the core. This means that the counts had been transformed into profiles of relative abundances following equation 8. Polynomial trend-surface and PCNM analyses will be used to detect structures in the multivariate diatom data within the core.

RDA of the multivariate diatom data from the RLGH core against level numbers showed that the core data contained a highly significant linear gradient ($R^2 = 0.190$, $R_a^2 = 0.182$, $p = 0.001$ after 999 random permutations; Figure 5a). We then analysed the response data against a 3rd-order polynomial of the core level numbers (1 to 101): the three monomials contributed significantly to the explanation of the diatom data, producing a model (not shown) with higher explanatory power ($R^2 = 0.460$, $R_a^2 = 0.443$, $p = 0.001$). All monomials of a 5th-order polynomial also contributed significantly to the explanation of the diatom data, producing a model with an even higher coefficient of determination ($R^2 = 0.567$, $R_a^2 = 0.545$, $p = 0.001$). Since the data seemed to be structured in an intricate series of scales, we turned to PCNM analysis to extract submodels corresponding to the different temporal scales present in the data.

The diatom data were regressed on level numbers to extract the linear gradient, as recommended by Borcard et al. (2004). PCNM analysis was then conducted on the detrended data, namely the residuals of these regressions. Sixty-eight PCNM variables were created using the 'PCNM' function of the PCNM R-language package (last paragraph of the previous section); these spatial variables, which have the form of sine waves of decreasing periods, represent variation at the various scales that can be identified in the series of 101 core levels. The first 50 PCNM variables, which had Moran's *I* coefficients larger than the expected value of *I* and thus modelled positive spatial correlation, were retained for canonical analysis. They were subjected to forward selection against the detrended diatom data, using the 'forward.sel' function of the PACKFOR package available at http://r-forge.r-project.org/R/?group_id=195; forward selection of explanatory variables in RDA is also available in the program CANOCO, version 4.5. Thirty PCNM variables were selected at the $\alpha = 0.05$ significance level (Monte Carlo permutation tests, 999 permutations). The selected PCNM variables were numbers 1 to 20, 28, 30, 32, 33, 35, 37, 38, 41, 42, and 45. Canonical redundancy analysis of the detrended diatom data by this subset of 30 PCNM variables explained R_a^2 = 70.1% of the variance in the detrended data. The RDA produced 9 significant canonical axes; three of them, which accounted for more than 5% of the detrended species variation, are displayed in Figure 5b-d. The diatom taxa contributing in an important way to the variation along these axes vary depending on the axis. Six species were highly positively correlated ($r > 0.6$) to the core level numbers (linear trend): *Tabellaria quadriseptata* (TA004A), *Navicula hoefleri* (NA167A), *Navicula cumbrensis* (NA158A), *Peronia fibula* (PE002A), *Eunotia denticulata* (EU015A), and *Eunotia naegelii* (EU048A); these species are found in the sections on the positive side of the linear trend (Figure 5a). Two taxa were highly negatively correlated ($r < –0.5$) to the same trend: *Brachysira brebissonii* (BR006A), *Cymbella* [PIRLA sp. 1] (CM9995). Two species were highly positively correlated to canonical axis 1 ($r > 0.5$, Figure 5b): *Aulacoseira perglabra* (AU010A), *Eunotia incisa* (EU047A); four taxa were highly negatively correlated ($r < –0.5$) to the same wave form: *Brachysira vitrea* (BR001A), *Achnanthes minutissima* (AC013A), *Tabellaria flocculosa* (TA001A), *Cymbellla perpusilla* (CM010A). And so on (Figures 5c-d). Each canonical axis displays structures representing a mixture of stratigraphical and temporal scales. This information could also be displayed in the form of biplots of the species together with the trend or with the PCNM variables.

Another useful way to describe the structure of the multivariate diatom data along the core is to separate the PCNM variables into an arbitrary number of groups, made of contiguous PCNMs, and examine the resulting submodels. We chose to divide them into three submodels. The broad-scale submodel contains PCNMs numbers 1 to 10 as explanatory variables; it explains R_a^2 = 47.7% of the detrended diatom variation. Canonical axes 1 to 3 of this fraction are significant and explain more than 5% of the detrended diatom variation (Figure 6a-c, $p = 0.001$). The taxa that are positively correlated with axis 1 ($r > 0.6$) are *Navicula krasskei* (NA044A) and *Aulacoseira perglabra* (AU010A); these species are found in the sections on the positive side of the wave form (Figure 6a). Other species that are highly negatively correlated with that axis ($r < –0.6$) are *Achnanthes linearis* (AC002A), *Tabellaria flocculosa* (TA001A), and *Eunotia iatriaensis* (EU019A); they are present in the sections found on the negative side of the wave form (Figure 6a). The medium-scale submodel uses PCNMs numbers 11 to 20 as explanatory variables; it explains R_a^2 = 9.1% of the detrended diatom variation. Only canonical axis 1 of that submodel is significant and explains more than 5% of the detrended diatom variation (Figure 6d). The fine-scale submodel uses PCNMs numbers 28, 30, 32, 33, 35, 37, 38, 41, 42, and 45 as explanatory variables. Taken alone, this submodel does not explain a significant portion of the diatom variation ($p = 0.916$, 999 permutations); the PCNM variables it contains were significant in the global model of 30 PCNMs, after the broad- and medium-scale PCNMs had been selected. None of its canonical axes is significant. We conclude that the core is mainly structured by processes operating at broad (5 x 10^3 - 10^4 year) and medium (5 x 10^2 - 10^3 year) scales. This PCNM example is presented simply to illustrate the potential of PCNM analysis in palaeolimnology. A more detailed analysis would naturally consider the estimated ages of each level (see Blaauw and Heegaard: Chapter 10 this volume) in the sediment core, rather than simply level numbers.

Testing hypotheses in (multi-)factorial experiments

RDA and CCA provide ways of testing hypotheses about multivariate data, as in analysis of variance (ANOVA). Assemblage composition data can be used as the response table in RDA provided that they are transformed in appropriate ways, as shown in Figure 3b-c. Examples of the use of RDA to test ANOVA-like hypotheses are found in Sabatier et al. (1989), ter Braak and Wiertz (1994), Verdonschot and ter Braak (1994), Legendre and Anderson (1999), and Hooper et al. (2002). The principle of this analysis is the following: multiple regression can be used to calculate any ANOVA model, provided that the factors are coded in appropriate ways in the matrix of predictors **X**. Since RDA and CCA are simply regression followed by PCA, they can be used in the same way as regression to carry out analysis of variance. The PCA portion of the procedure is only needed to illustrate the ANOVA results using bi- or triplots, as in Hooper et al. (2002); it is not needed nor computed for the test of significance of the canonical relationship. RDA and CCA use Monte Carlo permutation tests to assess the significance of the relationship between the response matrix **Y** (or **Q**) and the factor coded into matrix **X**.

Here are examples of such potential hypotheses in palaeolimnology. For sediment cores: in time series, are there differences between time periods of interest? In the comparison of cores: are there differences among cores which can be related to sampling regions? (In the latter example, one can control for the time pairing of core subunits by coding them into a matrix of covariables.) In such analyses, the factors (or ANOVA classification criteria) must be coded using dummy variables. A set of ordinary (binary 0-1) dummy variables will do the job when analysing a single factor. For two or more factors and their interactions, the factors must be coded using Helmert contrasts, also called orthogonal dummy variables. A method of coding such factors is described in Appendix C of Legendre and Anderson (1999) and in Legendre and Legendre (2012). In R, factors can be automatically coded into Helmert contrasts by function ***model.matrix()*** using an appropriate contrast type specification.

The use of RDA and CCA to test hypotheses in palaeolimnology is discussed in detail by Lotter and Anderson (Chapter 15 this volume). Birks (1996, 1998, 2010) reviewed hypothesis testing in palaeolimnology both directly through rigorous project design and site selection (e.g., Birks et al. 1990b) and indirectly through RDA or CCA.

Software

A list of programs and packages available for simple and canonical ordination of ecological and palaeoecological data is presented in Table 3. The list of functions available, especially in the R language (R Development Core Team 2011), is rapidly increasing.

Most general-purpose statistical programs contain procedures for principal component analysis (PCA). Very few allow, however, the direct drawing of biplots of species and objects, and many do not even compute the coordinates of the species and objects necessary to construct distance or correlation biplots. PCA and biplots are available in Canoco (biplots: CanoDraw), in PC-ORD, in Syn-Tax, and in the 'rda' function of vegan, the 'dudi.pca' function of ade4, and the 'pca' function of labdsv (R-language libraries).

Correspondence analysis (CA) is offered in few general-purpose statistical packages. In R, palaeoecologists will find it in the same packages as PCA. Principal coordinate analysis (PCoA) is available in the PrCoord program distributed with Canoco, in functions of the R language ('cmdscale' and its wrappers 'capscale', 'pco' and 'cmds.diss', in 'pcoa', and in 'dudi.pco'; see Table 3 for references), and in Syn-Tax. Non-metric multidimensional scaling (NMDS) is found in PC-ORD, in function metaMDS of the R language, and in isoMDS and its wrappers 'nmds' and 'bestnmds', and in Syn-Tax. NMDS is also found in some general-purpose statistical packages; they offer, however, a poor choice of dissimilarity functions.

Canoco and the 'rda' and 'cca' functions of the vegan R-language package are widely used for unconstrained or constrained ordination analysis. Other programs and packages allow the computation of some forms of canonical analysis: the PC-ORD and Syn-Tax packages, and the program Polynomial_RdaCca of Makarenkov and Legendre (2002). Canoco contains many interesting features for palaeoecologists, not shared by most other canonical analysis packages (Rejmánek and Klinger 2003), such as a procedure for selecting the environmental variables of **X** that contribute significantly to modelling **Y**; selection of explanatory variables is also available in the R language: functions 'ordistep' and 'ordiR2step' (vegan), as well as 'forward.sel' (packfor on http://r-forge.r-project.org/R/?group_id=195). Canoco also offers tests of significance for individual canonical eigenvalues (also in vegan), partial canonical analysis (also in vegan), and permutation methods especially designed for time series and blocked experimental designs.

Summary

The simple ordination methods mostly used by palaeoecologists and palaeolimnologists are *principal component analysis* (PCA) and *correspondence analysis* (CA), and, more rarely, *principal coordinate analysis* (PCoA) and *non-metric multidimensional scaling* (NMDS). These methods are reviewed in a geometric framework. They mostly differ by the types of distances among objects that they allow users to preserve during ordination. Canonical ordination methods are generalisations of the simple ordination techniques; the ordination is constrained to represent the part of the variation of a table of response variables (e.g., species abundances) that is maximally related to a set of explanatory variables (e.g., environmental variables). *Canonical redundancy analysis* (RDA) is the constrained form of PCA whereas *canonical correspondence analysis* (CCA) is the constrained form of CA. Canonical ordination methods have also been proposed that look for linear and polynomial relationships between the dependent and explanatory variables. Tests of significance can be obtained in canonical ordination, just as in multiple regression. Canonical ordination serves as the basis for variation partitioning, an analytical procedure widely used by palaeolimnologists.

Acknowledgements

We are indebted to Cathy Jenks for her invaluable help in the editing of this chapter. This is publication A210 for the Bjerknes Centre for Climate Research.

References

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman & Hall, London

Allott TEH, Harriman R, Battarbee RW (1992) Reversibility of lake acidification at The Round Loch of Glenhead, Galloway, Scotland. Envir Pollut 77 219-25

Ammann B, Birks HJB, Drescher-Schneider R, Juggins S, Lang G, Lotter A (1993) Patterns of variation in Late-glacial pollen stratigraphy along a northwest - southeast transect through Switzerland - a numerical analysis. Quat Sci Rev 12:277-286

Anderson MJ, Ellingsen KE, McArdle BH (2006) Multivariate dispersion as a measure of beta diversity. Ecology Letters 9:683-693

Anderson MJ, Legendre P (1999) An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. J Stat Comp Simul 62:271-303

Anderson NJ, Renberg I, Segerström U (1995) Diatom production responses to the development of early agriculture in a boreal forest lake-catchment (Kassjön, northern Sweden). J Ecol 83:809-822

Barker P, Telford RJ, Merdaci O, Williamson D, Taieb M, Vincens A, Gibert E (2000) The sensitivity of a Tanzanian crater lake to catastrophic tephra input and four millennia of climate change. Holocene 10:303-310

Birks HH, Birks HJB (2001) Recent ecosystem dynamics in nine North African lakes in the CASSARINA project. Aquat Ecol 35:461-478

Birks HH, Battarbee RW, Birks HJB (2000) The development of the aquatic ecosystem at Kråkenes Lake, western Norway, during the late-glacial and early-Holocene – a synthesis. J Paleolimnol 23:91-114

Birks HJB (1973) Modern pollen rain studies in some arctic and alpine environments. In: Birks HJB, West RG (eds), Quaternary Plant Ecology. Blackwell Scientific Publications, Oxford, pp.143-168

Birks HJB (1977) Modern pollen rain and vegetation of the St Elias Mountains, Yukon Territory. Can J Bot 55:2367-2382

Birks HJB (1985) Recent and possible mathematical developments in quantitative palaeoecology. Palaeogeogr Palaeoclim Palaeoecol 50:107-147

Birks HJB (1987) Multivariate analysis of stratigraphic data in geology: a review. Chemometrics Intell Lab Sys 2:109-126

Birks HJB (1992) Some reflections on the application of numerical methods in Quaternary palaeoecology. Publications of Karelian Institute 102 7-20

Birks HJB (1995) Quantitative palaeoenvironmental reconstructions. In: Maddy D, Brew S (eds), Statistical Modelling of Quaternary Science Data. Technical Guide 5, Quaternary Research Association, Cambridge, pp.161-254

Birks HJB (1986) Achievements, developments, and future challenges in quantitative Quaternary palaeoecology. INQUA Commission for the Study of the Holocene Sub-Commission on Data-Handling Methods Newsletter 14:1-8

Birks HJB (1998) Numerical tools in palaeolimnology – progress, potentialities, and problems. J Paleolimnol 20:307-332

Birks HJB (2008) Ordination – an ever-expanding tool-kit for ecologists? Bull Brit Ecol Soc 39:31-33

Birks HJB (2010) Numerical methods for the analysis of diatom assemblage data. In: JP Smol, EF Stoermer (eds), The Diatoms: Applications for the Environmental and Earth Sciences (2nd edition). Cambridge University Press, Cambridge, pp.23-54

Birks HJB This volume. Chapter 9 Stratigraphical data analysis. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Birks HJB, Berglund BE (1979) Holocene pollen stratigraphy of southern Sweden: a reappraisal using numerical methods. Boreas 8:257-279

Birks HJB, Gordon AD (1985) Numerical Methods in Quaternary Pollen Analysis. Academic Press, London

Birks HJB, Jones VJ This volume. Chapter 3 Data-sets. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Birks HJB, Lotter AF (1994) The impact of the Laacher See Volcano (11000 yr. BP) on terrestrial vegetation and diatoms. J Paleolimnol 11:313-322

Birks HJB, Peglar SM (1979) Interglacial pollen spectra at Fugla Ness, Shetland. New Phytol 68:777-796

Birks HJB, Juggins S, Line JM (1990a) Lake surface-water chemistry reconstructions from palaeolimnological data. In: Mason BJ (ed), The Surface Waters Acidification Programme. Cambridge University Press, Cambridge. pp.301-313

Birks HJB, Berge F, Boyle JF, Cumming BF (1990b) A palaeoecological test of the land-use hypothesis for recent lake acidification in south-west Norway using hill-top lakes. J Paleolimnol 4:69-85

Birks HJB, Austin HA, Indrevær NE, Peglar SM, Rygh C (1998) An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1996. Available from http://www.bio.umontreal.ca/Casgrain/cca_bib /index.html

Blaauw M, Heegaard E This volume. Chapter 10 Estimation of age-depth relationships. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Blanchet FG, Legendre P, Borcard D (2008) Modelling directional spatial processes in ecological data. Ecol Model 215:325-336

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol Model 153:51-68

Borcard D, Legendre P (2004) SpaceMaker2 – User's guide. Département de sciences biologiques, Université de Montréal. Program and user's guide available from http://www.bio.umontreal.ca/legendre/indexEn.html

Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. Ecology 73:1045-1055

Borcard D, Legendre P, Avois-Jacquet C, Tuomisto H (2004) Dissecting the spatial structure of ecological data at multiple scales. Ecology 85:1826-1832

Borcard D, Gillet F, Legendre P (2011) Numerical Ecology with R. Springer, New York.

Bradshaw EG, Rasmussen P, Odgaard BV (2005) Mid- to late-Holocene land-use change and lake development at Dallund Sø, Denmark: synthesis of multiproxy data, linking land and lake. Holocene 15:1152-1162

Brodersen KP, Odgaard BV, Vestergaard O, Anderson NJ (2001) Chironomid stratigraphy in the shallow and eutrophic Lake Søbygaard, Denmark: chironomid–macrophyte occurrence. Freshwater Biol 46:253-267

Brodersen KP, Whiteside MC, Lindegaard C (1998) Reconstruction of trophic state in Danish lakes using subfossil chydorid (Cladocera) assemblages. Can J Fish Aquat Sci 55:1093-1103

Burt C (1952) Tests of significance in factor analysis. Brit J Psychological Stat Section 5:109-133

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 21:550-570

Chessel D, Dufour AB, Thioulouse J (2004) The ADE4 package – I: One-table methods. R News 4:5-10

Dolédec S, Chessel D (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biol 31:277-294

Dray S, Chessel D, Thioulouse J (2003) Co-inertia analysis and the linking of ecological data tables. Ecology 84:3078-3089

Dray S, Dufour AB, Chessel D (2007) The ADE4 package – II: Two-table and K-table methods. R News 7:47-52

Dray S, Legendre P, Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol Model 196:483-493

Eastwood WJ, Tibby J, Roberts N, Birks HJB, Lamb HF (2002) The environmental impact of the Minoan eruption of Santorini (Thera): statistical analysis of palaeoecological data from Göhlisar, southwest Turkey. Holocene 12:431-444

Escofier B, Pagès, J (1994) Multiple factor analysis (AFMULT package). Computational Statistics and Data Analysis 18:121–140

Ezekiel, M. 1930. Methods of Correlation Analysis. Wiley & Sons, New York

Faith DP, Minchin PR, Belbin L (1987) Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57-68

Frontier S (1976) Étude de la décroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. J Experimental Mar Biol Ecol 25:67-75

Gasse F, Juggins S, Ben Khelifa L (1995) Diatom-based transfer functions for inferring past hydrochemical characteristics of African lakes. Palaeogeogr Palaeoclim Palaeoecol 117:31-54

Gauch HG (1982) Noise reduction by eigenvector ordination. Ecology 63 :1643-1649

Gordon AD (1982) Numerical methods in Quaternary palynology V. Simultaneous graphical representation of the levels and taxa in a pollen diagram. Rev Palaeobot Palynol 37:155-183

Gordon AD, Birks HJB (1974) Numerical methods in Quaternary palaeoecology. II. Comparisons of pollen diagrams. New Phytol 73:221-249

Gower JC (1968) Adding a point to vector diagrams in multivariate analysis. Biometrika 55:582-585

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857-871

Gower JC (1982) Euclidean distance geometry. Math Sci 7:1-14

Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. J Class 3:5-48

Griffith DA (2000) A linear regression solution to the spatial autocorrelation problem. J Geogr Sys 2:141-156

Griffith DA, Peres-Neto PR (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analysis. Ecology 87:2603-2613

Hall RI, Leavitt PR, Quinlan R, Dixit AS, Smol JP (1999) Effects of agriculture, urbanization, and climate on water quality in the northern Great Plains. Limnol Oceanogr 44:739-756

He F, Legendre P (1996) On species-area relations. Am Nat 148:719-737

He F, Legendre P (2002) Species diversity patterns derived from species-area models. Ecology 83:1185-1198

Hill MO (1974) Correspondence analysis: a neglected multivariate method. Appl Stat 23:340-354

Hill MO, Gauch HG (1980) Detrended correspondence analysis – an improved ordination technique. Vegetatio 42:47-58

Hooper E, Condit R, Legendre P (2002) Responses of 20 native tree species to reforestation strategies for abandoned farmland in Panama. Ecol Appl 12:1626-1641

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educat Psychol 24:417-441, 498-520

Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74:2204-2214

Jackson DA, Somers KM (1991) Putting things in order: the ups and downs of detrended correspondence analysis. Am Nat 137:704-712

Jacobson GL, Grimm EC (1986) A numerical analysis of Holocene forest and prairie vegetation in central Minnesota. Ecology 67:958-966

Jones VJ, Birks HJB (2004) Lake-sediment records of recent environmental change on Svalbard: results of diatom analysis. J Paleolimnol 31:445-466

Jones VJ, Juggins S (1995) The construction of a diatom-based chlorophyll *a* transfer function and its application at three lakes on Signy Island (maritime Antarctic) subject to differing degrees of nutrient enrichment. Freshwater Biol 34:433-445

Jones VJ, Juggins S, Ellis-Evans JC (1992) The relationship between water chemistry and surface sediment diatom assemblages in maritime Antarctic lakes. Antarct Sci 5:339-348

Jones VJ, Stevenson AC, Battarbee RW (1989) Acidification of lakes in Galloway, southwest Scotland: a diatom and pollen study of the post-glacial history of the Round Loch of Glenhead. J Ecol 77:1-23

Juggins S, Birks HJB This volume. Chapter 12 Quantitative environmental reconstructions from biological data. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Juggins S, Telford RJ This volume. Chapter 4 Exploratory data analysis and data display. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Kenkel NC, Orlóci L (1986) Applying metric and non-metric multidimensional scaling to ecological studies: some new results. Ecology 67:919-928

Korsman T, Segerström U (1998) Forest fire and lake-water acidity in a northern Swedish boreal area: Holocene changes in lake-water quality at Makkassjön. J Ecol 86:113-124

Korsman T, Renberg I, Anderson NJ (1994) A palaeolimnological test of the influence of Norway spruce (*Picea abies*) immigration on lake-water acidity. Holocene 4:132-140

Laliberté E, Shipley B (2010) FD: Measuring functional diversity from multiple traits, and other tools for functional ecology. R package version 1.0-7. http://cran.r-project.org/

Lamb HF (1984) Modern pollen spectra from Labrador and their use in reconstructing Holocene vegetational history. J Ecol 72:37-59

Leavitt PR, Findlay DL, Hall RI, Smol JP (1999). Algal responses to dissolved organic carbon loss and pH decline during whole-lake acidification: evidence from palaeolimnology. Limnol Oceanogr 44:757-773

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659-1673

Legendre P, Anderson MJ (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecol Monogr 69:1-24

Legendre P, Birks HJB This volume. Chapter 6 Clustering and partitioning. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Legendre P, Borcard D (2006) Quelles sont les échelles spatiales importantes dans un écosystème ? Chapter 19 in: Droesbeke J-J, Lejeune M, Saporta G (éds), Analyse Statistique de Données Spatiales. Éditions TECHNIP, Paris

Legendre P, Gallagher E (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129:271-280

Legendre P, Legendre L (1998) Numerical ecology (2nd English edition). Elsevier, Amsterdam

Legendre P, Legendre L (2012) Numerical ecology (3nd English edition). Elsevier, Amsterdam

Lepš J, Šmilauer P (2003) Multivariate Analysis of Ecological Data using CANOCO. Cambridge University Press, Cambridge

Lotter AF, Anderson NJ This volume. Chapter 15 Limnological response to environmental changes at inter-annual to decadal time-scales. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Lotter AF, Birks HJB (1993) The impact of the Laacher See Tephra on terrestrial and aquatic ecosystems in the Black Forest southern Germany. J Quat Sci 8:263-276

Lotter AF, Birks HJB (1997) The separation of the influence of nutrients and climate on the varve time-series of Baldegersee, Switzerland. Aquat Sci 59:362-375

Lotter AF, Birks HJB (2003) The Holocene palaeolimnology of Sägistalsee and its environmental history – a synthesis. J Paleolimnol 30:333-342

Lotter AF, Birks HJB, Hofmann W, Marchetto A (1997) Modern diatom, Cladocera, chironomid, and chrysophytes cyst assemblages as quantitative indicators for the reconstruction of past environmental conditions in the Alps. I. Climate. J Paleolimnol 18:395-420

Lotter AF, Birks HJB, Hofmann W, Marchetto A (1998) Modern diatom, Cladocera, chironomid, and chrysophytes cyst assemblages as quantitative indicators for the reconstruction of past environmental conditions in the Alps. II. Nutrients. J Paleolimnol 18:443-463

Lotter AF, Birks HJB, Zolitschka B (1995). Late-glacial pollen and diatom changes in response to two different environmental perturbations: volcanic eruption and Younger Dryas cooling. J Paleolimnol 14:23-47

Lotter AF, Eicher U, Birks HJB, Sigenthaler U (1992) Late-glacial climatic oscillations as recorded in Swiss lake sediments. J Quat Sci 7:187-204

MacDonald GM, Ritchie JC (1986) Modern pollen spectra from the Western Interior of Canada and the interpretation of late Quaternary vegetation development. New Phytol 103:245-268

Maechler M, Rousseeuw P, Struyf A, Hubert M (2005) Cluster analysis basics and extensions. R package version 1.12.1. http://cran.r-project.org/

Makarenkov V, Legendre P (2002) Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. Ecology 83:1146-1161

Minchin PR (1987). An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69:89-107

Motyka J (1947) O zadaniach i metodach badan geobotanicznych. Sur les buts et les méthodes des recherches géobotaniques. Annales Universitatis Mariae Curie-Sklodowska (Lublin, Polonia), Sectio C, Supplementum I.

Noy-Meir I, Walker D, Williams WT (1975) Data transformations in ecological ordination II. On the meaning of data standardization. J Ecol 63:779-800

Ochiai A (1957) Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bull Jpn Soc Sci Fish 22:526-530

Odgaard BV (1994) The Holocene vegetation history of northern West Jutland, Denmark. Opera Botanica 123:1-71

Odgaard BV, Rasmussen P (2000) Origin and temporal development of macro-scale vegetation patterns in the cultural landscape of Denmark. J Ecol 88:733-748

Ohtani K (2000) Bootstrapping R^2 and adjusted R^2 in regression analysis. Econom Model 17:473-483

Oksanen J, Blanchet B, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens MHH, Wagner H (2011) VEGAN: Community ecology package. R package version 1.17-0. http://cran.r-project.org/

Orlóci L (1967) An agglomerative method for classification of plant communities. J Ecol 55:193-205

Paradis E, Bolker B, Claude J, Cuong HS, Desper R, Durand B, Dutheil J, Gascuel O, Jobb G, Heibl C, Lawson D, Lefort V, Legendre O, Lemon J, Noel Y, Nylander J, Opgen-Rhein R, Strimmer K, de Vienne D (2010) ape: Analyses of phylogenetics and evolution. R package version 2.5. http://cran.r-project.org/

Peres-Neto, PR, Legendre P (2010) Estimating and controlling for spatial structure in the study of ecological communities. Global Ecol Biogeogr 19:174-184

Peres-Neto PR, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87: 2614-2625.

Pienitz R, Smol JP, Birks HJB (1995) Assessment of freshwater diatoms as quantitative indicators of past climatic change in the Yukon and Northwest Territories, Can J Paleolimnol 13:21-49

Pinel-Alloul B, Niyonsenga T, Legendre P (1995) Spatial and environmental components of freshwater zooplankton structure. Écoscience 2:1-19

Potapova MG, Charles DF (2002). Benthic diatoms in USA rivers; distributions along spatial and environmental gradients. J Biogeogr 29:167-187

Prentice IC (1977) Non-metric ordination methods in ecology. J Ecol 65:85-94

Prentice IC (1978) Modern pollen spectra from lake sediments in Finland and Finnmark, north Norway. Boreas 7:131-153

Prentice IC (1980) Multidimensional scaling as a research tool in Quaternary palynology: a review of theory and methods. Rev Palaeobot Palynol 31:71-104

Prentice IC (1986) Multivariate methods for data analysis. In: Berglund BE (ed) Handbook of Holocene Palaeoecology and Palaeohydrology. Wiley & Sons, Chichester, pp.775-797

Quinlan R, Leavitt PR, Dixit AS, Hall RI, Smol JP (2002) Landscape effects of climate, agriculture, and urbanization on benthic invertebrate communities of Canadian prairie lakes. Limnol Oceanogr 47:378-391

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rao CR (1995) A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió (Quaderns d'Estadística i Investigació Operativa) 19:23-63

Raup DM, Crick RE (1979) Measurement of faunal similarity in paleontology. J Paleontol 53:1213-1227

Rejmánek M, Klinger R (2003) CANOCO 4.5 and some comparisons with PC-ORD and SYN-TAX. Bull Ecol Soc Am 84:69-74

Renberg I, Korsman T, Birks HJB (1993) Prehistoric increases in the pH of acid-sensitive Swedish lakes caused by land-use changes. Nature 362:824-826

Renkonen O (1938) Statisch-ökologische Untersuchungen über die terrestiche Kaferwelt der finnischen Bruchmoore. Ann Zool Soc Bot Fenn Vanamo 6:1-231

Ritchie JC (1977) The modern and late Quaternary vegetation of the Campbell-Dolomite Uplands, near Inuvik, N.W.T., Canada. Ecol Monogr 47:401-423

Roberts DW (2007). LABDSV: Ordination and multivariate analysis for ecology. R package version 1.3-1. http://cran.r-project.org/, http://ecology.msu.montana.edu/labdsv/R

Sabatier R, Lebreton J-D, Chessel D (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In: Coppi R, Bolasso S (eds) Multiway Data Analysis. Elsevier, The Netherlands, pp.341-352

Seppä H (1996) Post-glacial dynamics of vegetation and tree-lines in the far north of Fennoscandia. Fennia 174:1-96

Simpson GL (2009) cocorresp: Co-correspondence analysis ordination methods. R package version 0.1-9. http://cran.r-project.org/

Simpson GL This volume. Chapter 13 Analogue methods in palaeolimnology. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Simpson GL, Hall RI This volume. Chapter 16 Human impacts – applications of numerical methods to evaluate surface-water acidification and eutrophication. In: Birks HJB, Lotter AF, Juggins S, Smol JP (eds) 2012. Tracking Environmental Change Using Lake Sediments Volume 5: Data Handling and Numerical Techniques. Springer, Dordrecht

Simpson GL, Shilland EM, Winterbottom JM, Keay J (2005) Defining reference conditions for acidified waters using a modern analogue approach. Environ Pollut 137:119-133

Sokal RR, Rohlf FJ (1995) Biometry – The Principle and Practice of Statistics in Biological Research (3rd edition). Freeman, New York

Telford RJ, Birks HJB (2005) The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. Quat Sci Rev 24:2173-2179

ter Braak CJF (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41:859-873

ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167-1179

ter Braak CJF (1987a) The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69:69-77

ter Braak CJF (1987b) Calibration. In: Jongman RHG, ter Braak CJF, van Tongeren OFR (eds) Data Analysis in Community and Landscape Ecology. Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge University Press, Cambridge, pp. 78-90.

ter Braak CJF (1987c) Ordination. In: Jongman RHG, ter Braak CJF, van Tongeren OFR (eds) Data Analysis in Community and Landscape Ecology. Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge University Press, Cambridge, pp.91-173

ter Braak CJF (1988a) Partial canonical correspondence analysis. In: Bock HH (ed) Classification and Related Methods of Data Analysis. North-Holland, Amsterdam, pp.551-558

ter Braak CJF (1988b) CANOCO – an extension of DECORANA to analyze species-environment relationships. Vegetatio 75:159-160

ter Braak CJF (1992) Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel K-H, Rothe G, Sendler W (eds) Bootstrapping and Related Techniques. Springer-Verlag, Berlin, pp.79-86

ter Braak CJF (1994) Canonical community ordination. Part I: Basic theory and linear methods. Écoscience 1:127-140

ter Braak CJF (1995) Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (k-nearest neighbours, partial least squares and weighted averaging partial least squares) and classical approaches. Chemometrics Intell Lab Syst 28:165-180

ter Braak CJF, Juggins S (1993) Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia 269:485-502

ter Braak CJF, Looman CWN (1987) Regression. In: Jongman RHG, ter Braak CJF, van Tongeren OFR (eds), Data Analysis in Community and Landscape Ecology. Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge University Press, Cambridge, pp.29-77

ter Braak CJF, Prentice IC (1988) A theory of gradient analysis. Adv Ecol Res 18:271-317

ter Braak CJF, Schaffers AP (2004) Co-correspondence analysis: a new ordination method to relate two community compositions. Ecology 85:834-846

ter Braak CJF, Šmilauer P (2002) CANOCO Reference Manual and CanoDraw User's Guide – Software for Canonical Community Ordination (version 4.5). Microcomputer Power, Ithaca, New York

ter Braak CJF, van Dam H (1989) Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia 178:209-223

ter Braak CJF, Verdonschot PFM (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquat Sci 57:255-289

ter Braak CJF, Wiertz J (1994) On the statistical analysis of vegetation change: a wetland affected by water extraction and soil acidification. J Veg Sci 5:361-372

Vellend M (2004) Parallel effects of land-use history on species diversity and genetic diversity of forest herbs. Ecology 85:3043-3055

Verdonschot PFM, ter Braak CJF (1994) An experimental manipulation of oligochaete communities in mesocosms treated with chlorpyrifos or nutrient additions: multivariate analyses with Monte Carlo permutation tests. Hydrobiologia 278:251-266

Vinebrooke RD, Hall RI, Leavitt PR, Cumming BF (1998) Fossil pigments as indicators of phototrophic response to salinity and climatic change in lakes of western Canada. Can J Fish Aquat Sci 55:668-681

Wartenberg D, Ferson S, Rohlf FJ (1987) Putting things in order: a critique of detrended correspondence analysis. Am Nat 129:434-448

Whittaker RH (1967) Gradient analysis of vegetation. Biol Rev (Camb.) 42:207-264

Table 1. Palaeolimnological uses of ordination analysis (abbreviations are explained below and in the main text).

Modern biological assemblages (e.g., diatoms, chironomids)

- Estimate the amount of compositional change or turnover – DCA
- Summarise graphically the major patterns of variation – PCA, tb-PCA, CA, DCA, more rarely PCoA or NMDS
- Display results of clustering or partitioning of data in a few dimensions – PCA, tb-PCA, CA, DCA, more rarely PCoA or NMDS

Modern environmental data (e.g., lake-water chemistry)

- Summarise graphically the major patterns of variation – PCA, more rarely PCoA or NMDS
- Display results of clustering or partitioning of data in a few dimensions – PCA, more rarely PCoA or NMDS

Fossil biological assemblages (e.g., diatoms, chironomids)

- Estimate the amount of compositional change or turnover – DCA or its canonical relative DCCA with object age or depth as the sole constraining variable
- Summarise graphically the major patterns of variation – PCA, tb-PCA, CA, DCA, more rarely PCoA or NMDS
- Summarise stratigraphically the major patterns of variation – plot PCA, tb-PCA, CA, or DCA ordination axis object scores (e.g., axes 1-3) stratigraphically
- Modelling temporal structure – RDA, tb-RDA, db-RDA, or CCA with PCNM temporal constraints

Down-core non-biological data (e.g., geochemistry, magnetics)

- Summarise graphically the major patterns of variation – PCA
- Summarise stratigraphically the major patterns of variation – plot PCA ordination axis object scores stratigraphically
- Modelling temporal structure – RDA with PCNM temporal constraints

Modern and fossil biological assemblages (e.g., diatoms, chironomids)

- Display similarities and dissimilarities between modern and fossil assemblages – PCA, tb-PCA, CA, DCA, more rarely PCoA or NMDS with either modern or fossil analysed passively or analysed together

Modern biological assemblages and modern environmental data (e.g., diatoms and lake-water chemistry)

- Estimate the amount of compositional change or turnover along individual environmental gradients – DCCA
- Summarise graphicaly the major patterns of biological variation explained by the environmental variables – RDA, tb-RDA, db-RDA, or CCA
- Summarise graphically the major patterns of biological variation remaining after the partialling of other environmental variables – partial RDA, partial tb-RDA, partial db-RDA, or partial CCA
- Assessment of statistical significance of single or combined environmental variables as predictors of the biological variation – RDA, tb-RDA, db-RDA, or CCA with Monte Carlo permutation tests
- Development of 'minimal adequate model' of environmental variables that explain statistically the biological variation almost as well as the full set of environmental varaiables – RDA, tb-RDA, db-RDA, or CCA with variable selection (e.g., forward selection)
- Partitioning biological variation among two or more sets of explanatory variables – RDA and partial RDA, tb-RDA and partial tb-RDA, db-RDA and partial db-RDA, CCA and partial CCA
- Modelling spatial structure – RDA, tb-RDA, db-RDA, or CCA with PCNM spatial constraints

Modern biological assemblages, modern environmental data, and fossil biological assemblages (e.g., diatoms and lake-water chemistry)

- Display similarities and dissimilarities between modern and fossil assemblages in relation to modern environmental gradients – RDA, tb-RDA, db-RDA, or CCA with the fossil assemblages analysed passively

Fossil biological assemblages and palaeoenvironmental variables (e.g., diatoms, occurrences of volcanic tephras)

- Test hypotheses of biological responses to particular environmental variables – RDA and partial RDA, tb-RDA and partial tb-RDA, db-RDA and partial db-RDA, CCA and partial CCA with Monte Carlo permutation tests
- Modelling temporal structure – PCNM

Fossil biological assemblages from many sites

- Summarise graphically the major patterns of variation – PCA, tb-PCA, CA, DCA, more rarely PCoA or NMDS
- Modelling spatial structure – RDA, tb-RDA, db-RDA, or CCA with PCNM spatial constraints

Abbreviations used in this table: CA, correspondence analysis; CCA: canonical correspondence analysis; db-RDA: distance-based canonical redundancy analysis; DCA: detrended correspondence analysis; DCCA, detetrended canonical correspondence analysis; NMDS, non-metric multidimensional scaling; PCA, principal component analysis; PCNM, principal coordinates of a neighbour matrix; PCoA, principal coordinate analysis; tb-PCA: transformation-based principal component analysis; tb-RDA, transformation-based canonical redundancy analysis.

Table 2. Questions that must be addressed prior to ordination analysis.

Transform physical data

- Univariate distributions are not symmetrical ⇒ Apply skewness-reduction transformation
- Variables are not in the same physical units ⇒ Apply standardisation or ranging
- Multistate qualitative variables ⇒ In some cases, transform them to dummy variables

Transform biological composition data (species presence-absence or abundance)

- Reduce asymmetry of distributions ⇒ Apply square root or log($y + c$) transformation
- Make biological composition data suitable for Euclidean-based ordination methods ⇒ Use the chord, chi-square, or Hellinger transformation

Choose an appropriate distance function

Popular similarity or distance functions are:

- Physical binary data: simple matching coefficient
- Species presence-absence data: Jaccard, Sørensen, and Ochiai coefficients. The transformation $D = \sqrt{1 - S}$ insures a fully Euclidean representation in principal coordinate analysis
- Quantitative physical data: Euclidean distance on standardised or ranged variables
- Physical data of mixed precision levels (quantitative, qualitative, binary): Gower similarity
- Species abundance data: the chord, chi-square, Hellinger coefficients, as well as Clark's coefficient of divergence, are Euclidean. The Steinhaus similarity (equivalent to the Odum/Bray-Curtis distance) and Whittaker's index of association may not be Euclidean

Table 3. Computer programs for ordination. The list makes no pretence at being exhaustive.

Simple ordination

- Canoco: PCA, CA
- PC-ORD: PCA, CA, NMDS
- PrCoord: PCoA (available dissimilarity measures: 7)
- R-language functions for PCA and CA: 'rda', 'cca' (vegan package); 'dudi.pca', 'dudi.coa' (ade4 package); 'pca' (labdsv package)
- R-language dissimilarity functions: 10 binary measures in 'dist.binary' of ade4, 6 dissimilarity measures in 'dist' of stats, 13 in 'vegdist' of vegan, 3 in 'daisy' of cluster, and 'gowdis' in FD
- R-language functions for PCoA: 'dudi.pco' (ade4 package); 'pcoa' (ape package); 'cmdscale' (stats package) and its wrappers 'cmds.diss' (mvpart package), 'pco' (labdsv package), and 'capscale' (vegan package).
- R-language functions for NMDS: 'isoMDS' (MASS package) and its wrappers 'nmds' and 'bestnmds' of labdsv, and 'metaMDS' of vegan. Dissimilarity measures available in the R language: see previous entry
- Syn-Tax: PCA, CA, PCoA, NMDS (available dissimilarity measures: 39)

Canonical ordination

- Canoco: linear RDA and CCA; partial RDA and CCA
- PC-ORD: linear CCA
- Polynomial_RdaCca: linear and polynomial RDA, linear and polynomial CCA
- R-language functions: 'rda' and 'cca' (vegan package): linear RDA and CCA; partial RDA and CCA
- R-language function for variation partitioning: 'varpart' (vegan package) partitions the variation of a response table **Y** with respect to two, three, or four explanatory tables **X**, using partial RDA
- Syn-Tax: linear RDA and CCA
- R-language package cocorresp: co-correspondence analysis

Biplots and triplots

- CanoDraw
- PC-ORD
- Syn-Tax
- R language: 'plot.cca' (vegan package) produces PCA and CA biplots as well as RDA and CCA triplots

Canoco, CanoDraw, and PrCoord (for Windows): available as a bundle from Scientia Publishing http://ramet.elte.hu/~scientia/ and Microcomputer Power http://www.microcomputerpower.com

PC-ORD (for Windows): available from MjM Software <http://home.centurytel.net/~mjm/pcordwin.htm>

R language (for Windows, Linux, and MacOS X): freely downloadable from the Comprehensive R Archive Network (CRAN) http://cran.r-project.org/. Libraries ade4 (Chessel et al. 2004, Dray et al. 2007), labdsv (Roberts 2007), cluster (Maechler et al. 2005), cocorresp (Simpson 2009), stats (R Development Core Team 2009), ape (Paradis et al. 2010), FD (Laliberté and Shipley 2010), and vegan (Oksanen et al. 2011)

Syn-Tax (for Windows and Macintosh): available from Scientia Publishing http://ramet.elte.hu/~scientia/ and Exeter Software http://www.exetersoftware.com

Polynomial_RdaCca (for Windows and Macintosh): freely downloadable from P. Legendre's Web page <http://www.bio.umontreal.ca/legendre/indexEn.html>

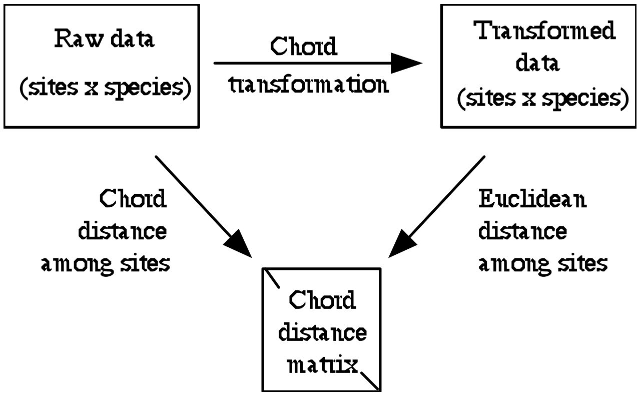

Figure 1. The role of the data transformations as a way of obtaining a given distance function. The example uses the chord distance. Modified from Legendre and Gallagher (2001).

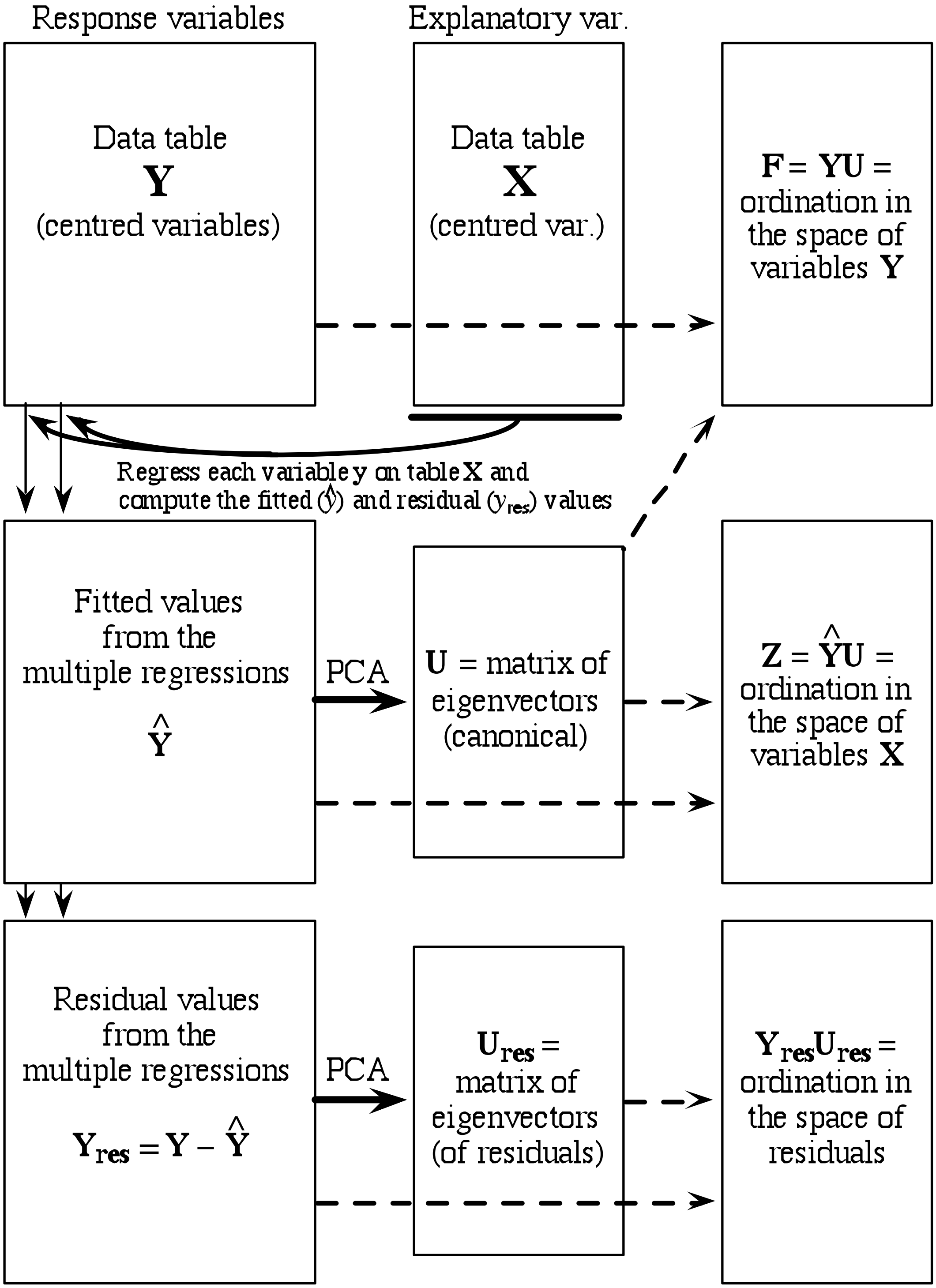

Figure 2. Redundancy analysis involves two steps: regression which produces fitted values $\hat{\mathbf{Y}}$ and residuals $\mathbf{Y}_{\text{res}}$, followed by PCA of the matrix of fitted values. PCA of the matrix of residuals may also be of interest. Modified from Legendre and Legendre (1998).

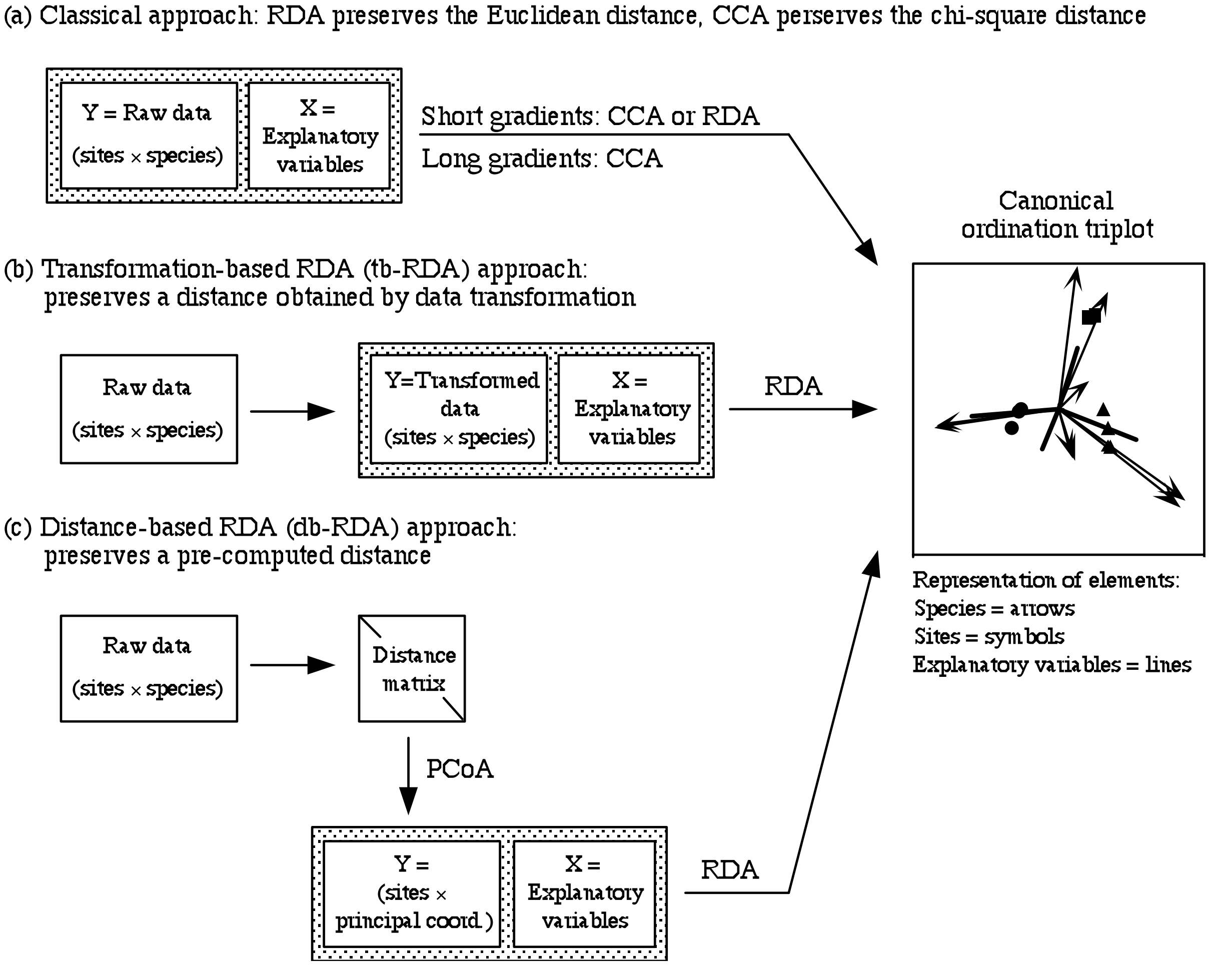
(a) Classical approach: RDA preserves the Euclidean distance, CCA perserves the chi-square distance
Y = Raw data
(sites × species)
X = Explanatory variables
Short gradients: CCA or RDA
Long gradients: CCA
Canonical ordination triplot
(b) Transformation-based RDA (tb-RDA) approach: preserves a distance obtained by data transformation
Raw data
(sites × species)
Y=Transformed data
(sites × species)
X = Explanatory variables
RDA
(c) Distance-based RDA (db-RDA) approach: preserves a pre-computed distance
Raw data
(sites × species)
Distance matrix
PCoA
Y = (sites × principal coord.)
X = Explanatory variables
RDA
Representation of elements:
Species = arrows
Sites = symbols
Explanatory variables = lines

Figure 3. Comparison of (a) classical RDA and CCA to (b,c) alternative approaches forcing canonical analyses to preserve other distances adapted to assemblage composition data. Modified from Legendre and Gallagher (2001).

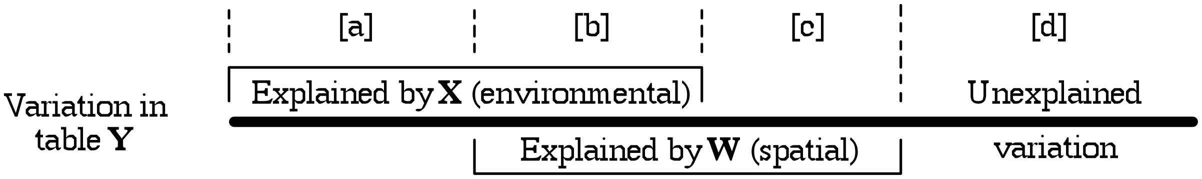
[a]
[b]
[c]
[d]
Variation in table **Y**
Explained by **X** (environmental)
Explained by **W** (spatial)
Unexplained variation

Figure 4. Partitioning the variation of a response data table **Y** with respect to a table **X** of environmental variables and a table **W** of spatial variables. The length of the thick horizontal line represents the total variation in **Y**. Modified from Borcard et al. (1992) and Legendre (1993).

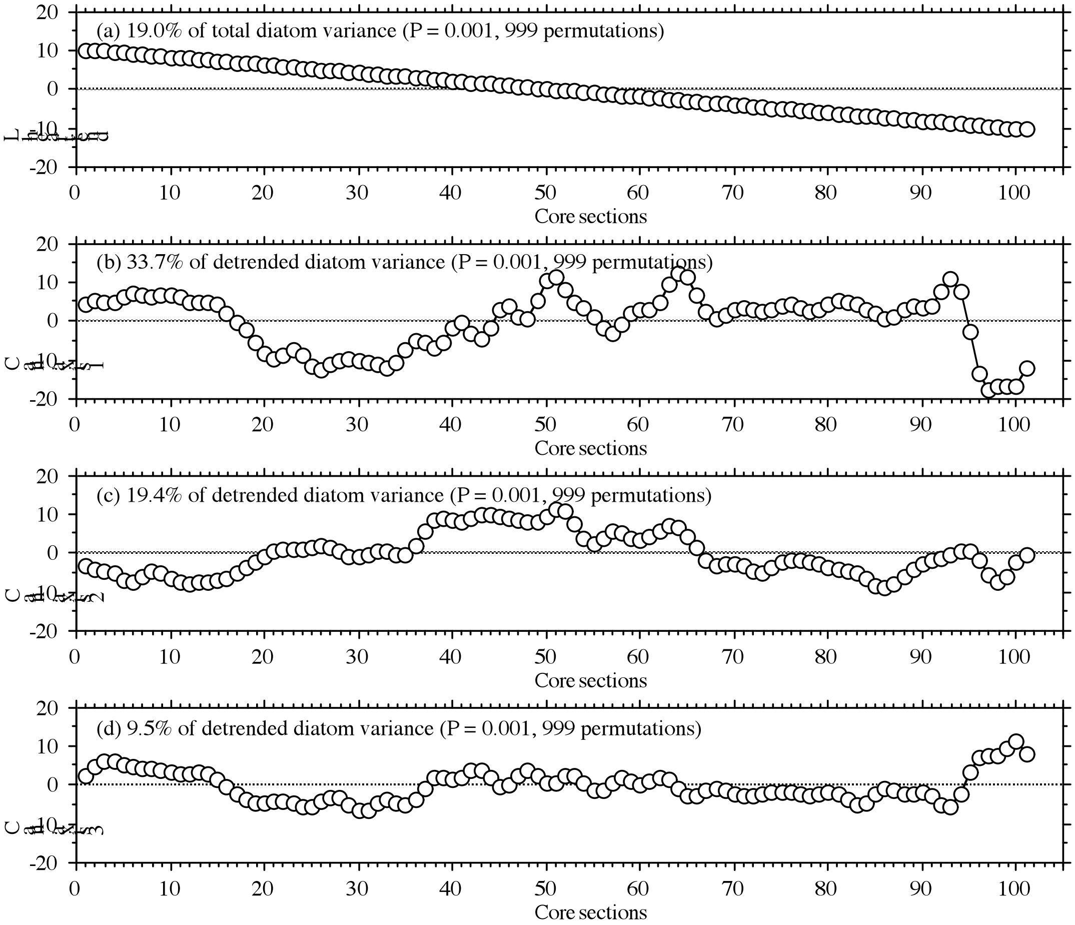

Figure 5. The linear gradient (a) and first three canonical axes of the PCNM model (b-d), as a function of the core level or section numbers in The Round Loch of Glenhead diatom stratigraphical data.

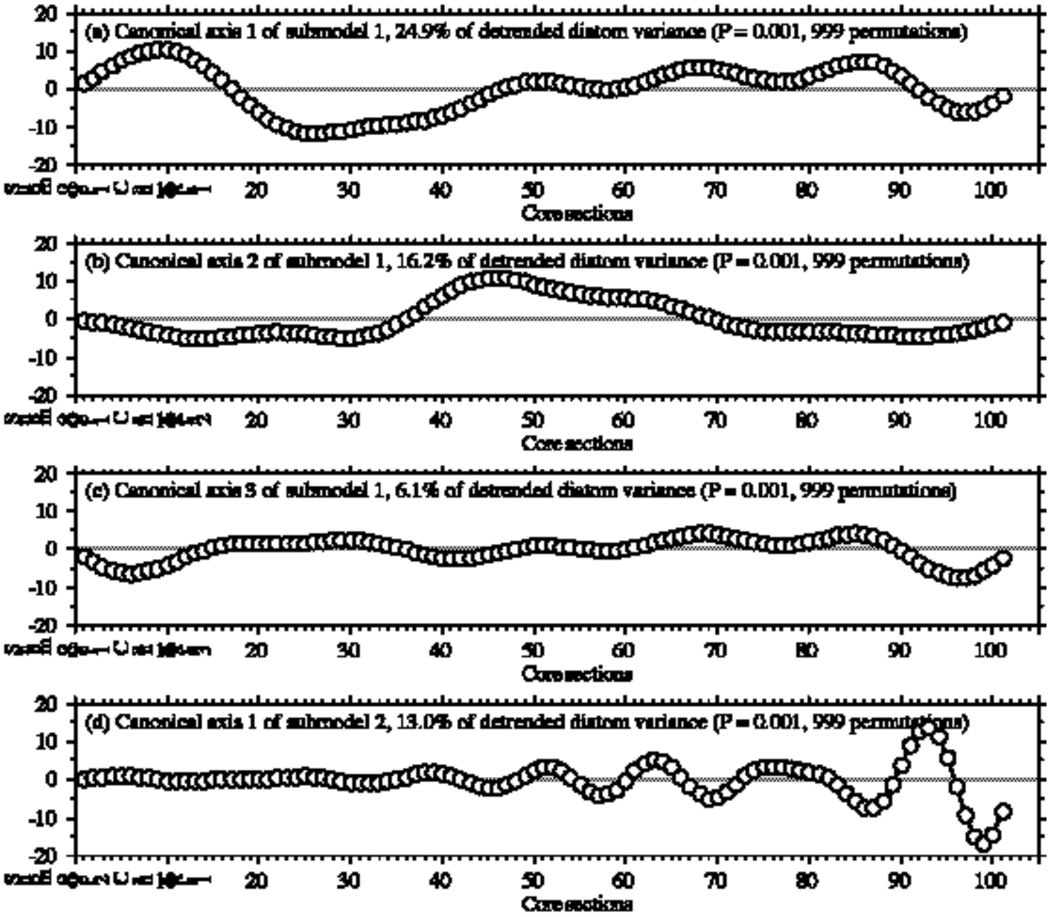

Figure 6. Significant canonical axes of the first two PCNM submodels, as a function of the core level or section numbers in The Round Loch of Glenhead diatom stratigraphical data.

.