

CHAPTER

9

QUANTITATIVE PALAEOENVIRONMENTAL
RECONSTRUCTIONS FROM HOLOCENE
BIOLOGICAL DATA

H.J.B. Birks

Abstract: There are three main approaches to reconstructing quantitatively Holocene environments from fossil biological assemblages: (1) indicator species, (2) assemblage or modern analogues (including response surfaces) and (3) multivariate transfer functions. Their assumptions, strengths and limitations are discussed. The basic concepts, requirements and numerical procedures involved in transfer functions are reviewed, and the properties of weighted averaging and weighted averaging partial least squares regression are outlined. A reconstruction of Holocene mean July temperatures in northern Norway from pollen-stratigraphical data is presented. Possible future developments are discussed.

Keywords: Climate reconstruction, Error estimation, Transfer functions, Weighted averaging, Weighted averaging partial least squares

Many Holocene palaeoecological studies aim to reconstruct features of the past environment from fossil assemblages preserved in sediments. Although fossil assemblages are usually studied quantitatively with individual pollen, chironomids, etc. being identified and counted, environmental reconstructions may be qualitative and presented as 'cool', 'warm', etc. The need for quantification in Holocene research is increasing, largely in response to demands for quantitative reconstructions of past environments as input to, or validation of, simulations by Earth System models of past, present and future climate patterns. There are three main approaches to reconstructing quantitatively past environments from biostratigraphical data (Birks, 1981, 1995, 1998; Birks and Birks, 1980): (1) indicator species approach, (2) assemblage approach and (3) multivariate indicator species approach involving mathematical transfer functions. All require information about modern environmental requirements of the taxa found as fossils. The basic assumption is methodological uniformitarianism (Rymer, 1978; Birks and Birks, 1980), namely that modern-day observations and relationships can be used as a model for past conditions and, more specifically, that organism–environment relationships have not changed with time, at least in the Late Quaternary.

This chapter discusses the multivariate indicator species approach involving transfer functions as a means of quantitatively reconstructing Holocene climates from fossil assemblages. It presents the basic concepts of transfer functions, the assumptions, requirements and data properties of the approach, the relevant numerical procedures, an application, and a discussion of the limitations of the approach. In order to put this approach into its palaeoecological context, I briefly discuss the indicator species and assemblage approaches. I conclude by suggesting

possible areas for future development. I draw extensively on reviews by Birks (1981, 1995, 1998) and ter Braak (1995).

9.1 INDICATOR SPECIES APPROACH

Fossil occurrences of a species with known modern environmental tolerances provide a basis for environmental reconstructions. Assuming methodological uniformitarianism, the past environment is inferred to have been within the environmental range occupied by the species today. This approach requires information about what environmental factors influence the distribution and abundance today of the species concerned. The commonest means of obtaining such information is to compare present-day distributions of species with selected climatic variables of potential ecological and physiological significance, such as mean temperature of the coldest month or maximum summer temperature (Dahl, 1998). If the geographical trend of an eco-climatic variable covaries with the species distribution in question, a cause-and-effect relationship is often assumed. For example, Conolly and Dahl (1970) related the modern distribution of *Betula nana* in the British Isles to the 22 °C maximum summer temperature isotherm for the highest points in areas where *B. nana* grows today. Fossil records show that it occurred widely in lowland Britain during the Late Glacial, where maximum summer temperatures are 30 °C today. Conolly and Dahl proposed therefore that there was a depression of 8 °C in maximum summer temperatures in the Late Glacial.

In some instances it is more realistic to consider species distributions in relation to two or more variables (Hintikka, 1963). This 'bivariate' approach was pioneered by Iversen (1944) in his classic work on *Viscum album*, *Hedera helix* and *Ilex aquifolium*. On the basis of detailed observations, Iversen delimited the 'thermal limits' within which they flowered and produced seed. He showed that *Ilex* is intolerant of cold winters but tolerant of cool summers, *Hedera* is intolerant of winters with mean temperatures colder than -1.5 °C but requires warmer summers than *Ilex*, and *Viscum* is tolerant of cold winters but requires warmer summers than either *Ilex* or *Hedera*. These shrubs are ideal indicator species because their pollen is readily identifiable to species level, it is not blown great distances so interpretative problems arising from far-distance transport rarely arise, and their berries are rapidly dispersed by birds. Their distributions are likely to be in equilibrium with climate. From fossil pollen occurrences, Iversen used this approach to suggest that Mid-Holocene summers were 2–3 °C warmer and winters 1–2 °C warmer than today in Denmark.

This approach has been extended to several species simultaneously to identify areas of climatic overlap for pollen (e.g. Grichuk, 1969; Markgraf *et al.*, 1986; McKenzie and Busby, 1992; Pross *et al.*, 2000), plant macrofossils (Sinka and Atkinson, 1999), chironomids (Dimitriadis and Cranston, 2001), molluscs (Moine *et al.*, 2002) and beetles (Elias, 1997; Atkinson *et al.*, 1987), the so-called mutual climatic range approach. This approach assumes that spatial correspondence between species distributions and selected climatic variables implies a causal relationship. As discussed by Birks (1981) and Huntley (2001), it can be unwise to assume such relationships exist. When climatic factors have been studied in detail (e.g. Forman, 1964; Pigott, 1970, 1981, 1992; Pigott and Huntley, 1991) many factors may be operative at different spatial scales. Moreover, only a few indicator taxa are usually considered and little or no attention is given to the numerical frequencies of the different taxa in the fossil assemblages. An alternative approach, considering the composition and abundance of the whole assemblage, is the assemblage approach.

9.2 ASSEMBLAGE APPROACH

This considers the fossil assemblage as a whole and the proportions of its different fossil taxa. It has been widely used in an intuitive non-quantitative way for many decades. For example, pollen assemblages are interpreted as reflecting tundra, pine forest, or deciduous forest. Past environmental inferences are based on the present-day environment in which these vegetation types occur. More recently it has been put on a more quantitative basis, the so-called **modern analogue technique** (MAT) and related response surface approach.

The basic idea of MAT is to compare numerically, using a dissimilarity measure (e.g. squared chord distance) (Overpeck *et al.*, 1985), the fossil assemblage with modern assemblages. Having found the modern sample(s) that is most similar to the fossil sample, the past environment for that sample is inferred to be the modern environmental variable(s) for the analogous modern sample(s). The procedure is repeated for all fossil samples and a simultaneous reconstruction of several environmental variables can be made using modern analogues. The environmental reconstruction(s) can be based on the modern sample that most closely resembles the fossil assemblage or, more reliably, it can be based on a mean or weighted mean of, say, the 10 or 25 most similar modern samples, with weights being the inverse of the dissimilarities so that modern samples with the lowest dissimilarity have the greatest weight in the reconstruction. Examples of MAT for reconstructing Holocene climates include Bartlein and Whitlock (1993) and Cheddadi *et al.* (1998). MAT has been extended by Guiot (1990, Guiot *et al.*, 1992, 1993a) to reconstruct several climatic variables from pollen assemblages for the last interglacial–glacial cycle.

Response surfaces are three-dimensional graphical representations of the occurrence and/or abundance of individual taxa in modern environmental space (Huntley, 1993). The *x* and *y* axes represent environmental variables and the *z* axis represents the occurrence or relative abundance of the taxon of interest. Modern pollen–climate response surfaces have been constructed to illustrate relative abundances of modern pollen varying along major climatic gradients (e.g. Bartlein *et al.*, 1986). The surfaces are fitted by multiple regression (Bartlein *et al.*, 1986) or locally weighted regression (Bartlein and Whitlock, 1993). Palaeoenvironmental reconstructions from Holocene assemblages (e.g. Allen *et al.*, 2002) are made by ‘stacking’ modern surfaces to produce synthetic assemblages for a series of grid nodes, usually 20×20 nodes, in modern climate space. These synthetic assemblages are then compared to fossil assemblages by a dissimilarity measure, usually the squared chord distance as in MAT. Climate values for the 10 grid nodes with synthetic pollen spectra most similar to the fossil assemblages are used to infer the past climate. The final inferred value is a mean of the climate values weighted by the inverse of the squared chord distances (Prentice *et al.*, 1991). Environmental reconstructions are thus done by MAT but the modern data consist of fitted pollen values in relation to modern climate and not the original pollen values. The fitted values naturally smooth the data to varying degrees depending on the smoothing procedures used (ter Braak, 1995). Much inherent local site variability that is assumed to be unrelated to broad-scale climate is removed (Bartlein and Whitlock, 1993).

Limitations in using MAT and response surfaces are the need for high-quality, taxonomically consistent modern data sets from comparable sedimentary environments as the fossil data and covering a wide environmental range, the absence of any reliable error estimates for

reconstructed values, the problems of defining 'good', 'poor' and 'no analogues', selecting an appropriate dissimilarity measure and the occurrence of no analogues and multiple analogues (Birks, 1995). No analogues arise when no modern assemblages are similar to the fossil assemblage (Huntley, 1996). Multiple analogues arise when the fossil assemblage is similar to several modern samples that differ widely in climate (Huntley, 1996, 2001), for example assemblages dominated by pine pollen that can be derived from northern, central, or Mediterranean Europe, all of which have very different climates today. Constraints can be built into the analogue-matching procedure to help minimize the multiple analogue problem by restricting possible analogues to be from the same modern biome as the biome reconstructed from the fossil assemblage (Huntley, 1993; Allen *et al.*, 2000).

The multivariate indicator species approach involving transfer functions attempts to overcome some of these problems.

9.3 MULTIVARIATE TRANSFER FUNCTION APPROACH

The idea of quantitative environmental reconstructions involving transfer functions is summarized in Fig. 9.1. There is one or more environmental variable X_0 to be reconstructed from fossil assemblages Y_0 consisting of m taxa in t samples. To estimate X_0 , we model the responses of the same m taxa today in relation to the environmental variable(s) (X). This involves a modern 'training set' or 'calibration set' of m taxa at n sites (Y) studied as assemblages preserved in surface sediments (e.g. surficial lake muds, ocean sediments), with associated modern environmental variables (X) for the same n sites. The modern relationships between Y and X are modelled numerically and the resulting transfer function is used to transform the fossil data Y_0 into quantitative estimates of the past environmental variable(s) (X_0). The various stages are schematically shown in Fig. 9.2.

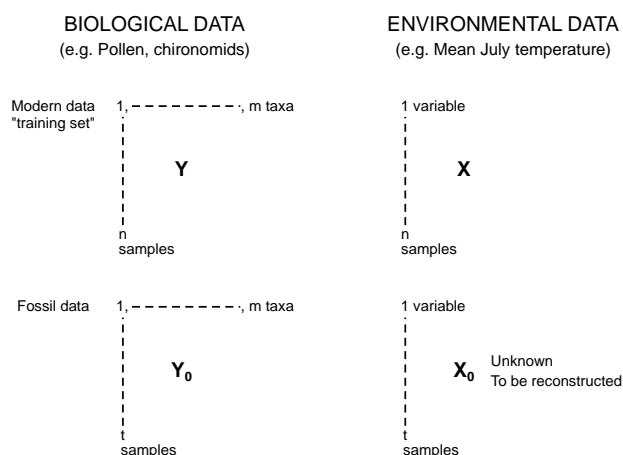


Figure 9.1 The principles of quantitative palaeoenvironmental reconstruction showing X_0 , the unknown environment variable (e.g. July temperature) to be reconstructed from fossil assemblages Y_0 , and the essential role of a modern training set consisting of modern biological (Y) and environmental (X) data.

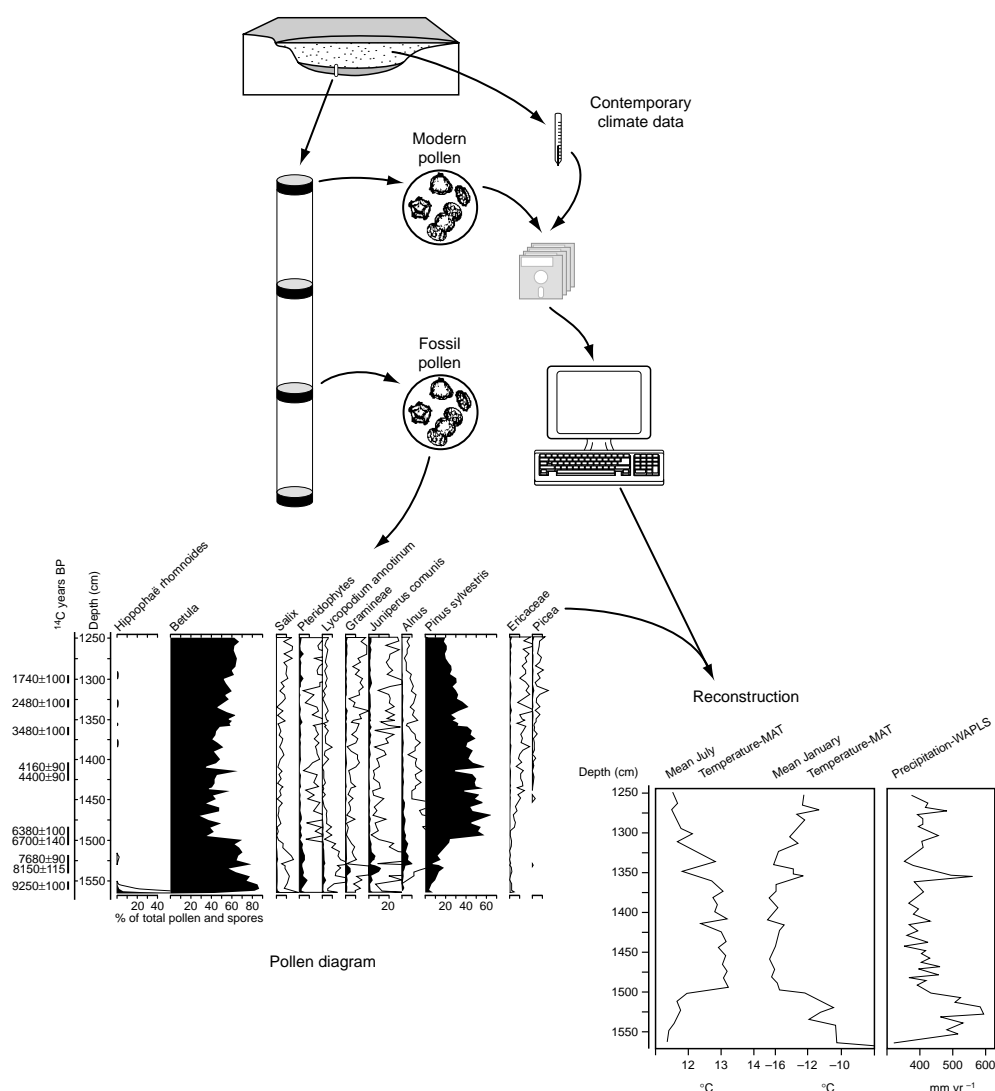


Figure 9.2 A schematic representation of the stages involved in deriving a quantitative reconstruction of past climate from pollen-stratigraphical data using a modern calibration or training set. Modified from an unpublished diagram by Steve Juggins.

Since Imbrie and Kipp (1971) revolutionized Quaternary palaeoecology by presenting, for the first time, a numerical procedure for quantitatively reconstructing past environments from fossil assemblages, several numerical techniques have been developed for deriving transfer functions (Birks, 1995). Some have a stronger theoretical basis, either statistically, ecologically, or both, than others. Some (e.g. weighted averaging partial least squares (WA-PLS) and its simpler relative, two-way weighted averaging (WA) regression and calibration) fulfil the basic requirements for quantitative reconstructions, perform consistently well with a range of data, do not involve an excessive number of parameters to be estimated and fitted and are thus relatively robust statistically and computationally economical.

As ter Braak (1995, 1996) and Birks (1995, 1998) discuss, there is a major distinction between models assuming a linear or monotonic response and a unimodal response between organisms and their environment, and between classical and inverse approaches for deriving transfer functions. It is a general law of nature that organism–environment relationships are usually non-linear and taxon abundance is often a unimodal function of the environmental variables. Each taxon grows best at a particular optimal value of an environmental variable and cannot survive where the value of that variable is too low or too high (ter Braak, 1996). Thus all taxa tend to occur over a characteristic but limited environmental range and within this range to be most abundant near their environmental optimum. The distinction between classical and inverse models is less clear (ter Braak, 1995).

In the classical approach taxon responses (Y) are modelled as a function of the environment (X) with some error:

$$Y = f(X) + \text{error}$$

The function $f()$ is estimated by linear, non-linear and/or multivariate regression from the modern training set. Estimated $f()$ is then 'inverted' to infer the unknown past environment from Y_0 . 'Inversion' involves finding the past environmental value that maximizes the likelihood of observing the fossil assemblage in that environment. If function $f()$ is non-linear, which it almost always is, non-linear optimization procedures are required (e.g. Birks *et al.*, 1990; Line *et al.*, 1994). Such procedures are not without programming problems (e.g. Birks, 2001b) and can be computationally demanding (e.g. Vasko *et al.*, 2000; Toivonen *et al.*, 2001).

In the inverse approach this difficult inversion is avoided by estimating directly the function (g) from the training set by inverse regression of X on Y :

$$X = g(Y) + \text{error}$$

The inferred past environment (X_0), given fossil assemblage (Y_0) is simply the estimate

$$X_0 = g(Y_0)$$

As ter Braak (1995) discusses, statisticians have debated the relative merits of both approaches. Inverse models perform best if the fossil assemblages are similar in composition to samples in the central part of the modern data, whereas classical models may be better at the extremes and under some extrapolation, as in 'no-analogue' situations. In the few comparisons of these two major approaches, inverse models (e.g. WA or WA-PLS) nearly always perform as well as classical models of Gaussian or multinomial logit regression in a non-Bayesian (ter Braak *et al.*, 1993; ter Braak, 1995; Birks 1998) or a Bayesian framework (Vasko *et al.*, 2000; Toivonen *et al.*, 2001) but with a fraction of the computing resources of classical approaches.

I will only consider WA and WA-PLS, as they currently represent simple, robust approaches for quantitative reconstructions. Ter Braak *et al.* (1993) concluded 'until such time that such sophisticated methods mature and demonstrate their power for species–environment calibration, WA-PLS is recommended as a simple and robust alternative.'

9.3.1 Assumptions

There are five major assumptions in quantitative palaeoenvironmental reconstructions (Imbrie and Webb, 1981; Birks *et al.*, 1990).

1. The m taxa in the modern data (\mathbf{Y}) are systematically related to the environment (\mathbf{X}) in which they live.
2. The environmental variable(s) to be reconstructed is, or is linearly related to, an ecologically important determinant in the system of interest.
3. The m taxa in the training set (\mathbf{Y}) are the same biological entities as in the fossil data (\mathbf{Y}_0) and their ecological responses to the environmental variable(s) of interest have not changed over the time represented by the fossil assemblage. Contemporary spatial patterns of taxon abundance in relation to \mathbf{X} can be used to reconstruct changes in \mathbf{X} through time.
4. The mathematical methods adequately model the biological responses to the environmental variable(s) of interest and yield transfer functions with sufficient predictive power to allow accurate and unbiased reconstructions of \mathbf{X} .
5. Other environmental variables than the ones of interest have negligible influence, or their joint distribution with the environmental variable in the past is the same as today.

9.3.2 Data Properties

Modern training sets (e.g. pollen, diatoms, etc.) contain many taxa (e.g. 50–300), whereas there may be 50–200 samples. Data are usually quantitative and commonly expressed as percentages of the total sample count. They are thus closed, multivariate compositional data with a constant-sum constraint. They often contain many zero values (up to 75 per cent of all entries) for sites where taxa are absent. The data are complex, showing noise, redundancy and internal correlations, and often contain outliers. Taxon abundance is usually a unimodal function of the environmental variables.

Modern environmental data usually contain fewer variables (*c.* 1–10) than the corresponding biological data. Environmental data rarely contain zero values. Quantitative environmental variables often follow a log-normal distribution and commonly show linear relationships and high correlations between variables (e.g. mean July temperature, number of growing-day degrees). There is thus often data redundancy.

9.3.3 Requirements

There are nine major requirements for quantitative reconstructions (Birks, 1995).

1. A biological system is required that produces abundant identifiable and preservable fossils and is responsive to the environmental variable(s) of interest today at the spatial and temporal scales of study.
2. A large high-quality training set is available. This should be representative of the likely range of past environmental variables, have consistent taxonomy and nomenclature and be of comparable quality (counting techniques, size, sampling methodology, preparation procedures, etc.) and from the same sedimentary environment (e.g. lakes).
3. Fossil data sets used for reconstruction should be of comparable taxonomy, quality and sedimentary environment as the training set.
4. Good independent chronology is required for the fossil data sets to permit correlations, comparisons and, if required, assessments of rates of biotic and environmental change.
5. Robust statistical models are required that can model the non-linear relationships between modern taxa and their environment and take account of the numerical properties of the biological data.

6. Reliable estimation of prediction errors is required. As the reliability of the reconstructed environmental values can vary from sample to sample, depending on composition, preservation, etc., sample-specific prediction errors are needed.
7. Critical evaluations and validations of all reconstructions are essential as any statistical procedure will produce a result. What matters is whether the result is ecologically sensible and statistically reliable.
8. The numerical methods are theoretically sound statistically and ecologically, easy to understand, robust, perform well with large and small data sets and taxon-poor and taxon-rich assemblages and are not too demanding in terms of computer resources.
9. The relevant computer programs are available to the research community.

9.3.3.1 Two-way Weighted Averaging (WA)

The basic idea behind WA (ter Braak, 1996) is that at a site with a particular environmental variable x , taxa with optima for x close to the site's value of x will tend to be the most abundant taxa present, if the taxa show a unimodal relationship with x . A simple and ecologically reasonable estimate of a taxon's optimum for x is the average of all the x values for sites at which the taxon occurs, weighted by the taxon's relative abundance. The estimated optimum is the weighted average of x . Taxon absences have no weight. The taxon's tolerance can be estimated as the weighted standard deviation of x . An estimate of a site's value of x is the weighted average of the optima for x for all the taxa present. Taxa with a narrow tolerance for x can, if required, be given greater weight than taxa with a wide tolerance. The underlying theory of WA and the conditions under which it approximates Gaussian logit regression and calibration are discussed by ter Braak (1996), ter Braak and Looman (1986) and ter Braak and Barendregt (1986).

Because the computations involved in WA are simple and fast, computer-intensive bootstrapping (Efron and Tibshirani, 1993) can be used to estimate the root mean square error of prediction (RMSEP) for inferred values of x for all modern samples, the whole training set and individual fossil samples (Birks *et al.*, 1990; Line *et al.*, 1994). The idea of **bootstrap error estimation** is to do many bootstrap cycles, say 1000. In each, a subset of modern samples is selected randomly but with replacement from the training set to form a bootstrap set of the same size as the original training set. This mimics sampling variation in the training set. As sampling is with replacement, some samples may be selected more than once in a cycle. Any modern samples not selected form a bootstrap *test* set for that cycle. WA is then used with the bootstrap training set to infer the variable of interest for the modern samples (all with known observed modern values) in the bootstrap *test* set. In each cycle, WA is also used to infer the environmental variable, x_0 , for each fossil sample. The standard deviation of the inferred values for both modern and fossil samples is calculated. This comprises one component of the overall prediction error, namely estimation error for the taxon parameters. The second component, due to variations in taxon abundance at a given environmental value, is estimated from the training set by the root mean square of the difference between observed values of x and the mean bootstrap of x when the modern sample is in the bootstrap *test* set. The first component varies from fossil sample to fossil sample, depending on the composition of the fossil assemblage, whereas the second component is constant for all fossil samples. The estimated RMSEP for a fossil sample is the square root of the sum of squares for these two components (Birks *et al.*, 1990).

WA has gained considerable popularity in palaeoecology in the last decade for various reasons.

1. It combines ecological realism (unimodal species responses and species-packing model) with mathematical and computational simplicity, rigorous underlying theory and good empirical power.
2. It does not assume linear species–environment responses, it is relatively insensitive to outliers and it is not hindered by multicollinearity between variables or by the large number of taxa in training sets.
3. Because of WA's computational simplicity, it is possible to use bootstrapping to derive RMSEP for all samples.
4. WA performs well in 'no-analogue' situations (Hutson, 1977; ter Braak *et al.*, 1993). In such situations, environmental inferences are based on the WA of the optima of taxa in common between the modern and fossil assemblages. As long as there are reliable optima estimates for the fossil taxa of high numerical importance, WA inferences are often relatively realistic. WA is thus a multivariate indicator species approach rather than an analogue-matching procedure.
5. WA appears to perform best with noisy, species-rich compositional data with many taxa absent from many samples and extending over a relatively long environmental gradient. WA does, however, have two important weaknesses (ter Braak and Juggins, 1993).
 - a. WA is sensitive to site distribution within the training set along the environmental gradient of interest (ter Braak and Looman, 1986).
 - b. WA disregards residual correlations in the biological data, namely correlations that remain in the biological data after fitting the environmental variable of interest that result from environmental variables not considered directly in WA. The incorporation of partial least squares (PLS) regression (Martens and Næs, 1989) into WA (ter Braak and Juggins, 1993) helps overcome the second weakness by utilizing residual correlations to improve optima estimates.

9.3.3.2 Weighted Averaging Partial Least Squares Regression (WA-PLS)

The relevant feature of PLS is that components are selected not to maximize the variance of each component within **Y** as in principal components analysis but to maximize covariance between components that are linear combinations of the variables within **Y** and **X**. In the unimodal equivalent, WA-PLS, components are selected to maximize covariance between the vector of weighted averages of **Y** and **X**. Subsequent components are chosen to maximize the same criterion but with the restriction that they are orthogonal and hence uncorrelated to earlier components (ter Braak *et al.*, 1993). Ter Braak and Juggins (1993) show that, with a small modification, WA is equivalent to the first PLS component of suitably transformed data. In WA-PLS, further orthogonal components are obtained as WA of the residuals for the environmental variable, in other words the regression residuals of *x* on the components extracted to date are used as new sample scores in the basic WA algorithm. A joint estimate of *x* is, in PLS, a linear combination of the WA-PLS components, each of which is a WA of the taxon scores, hence the name WA-PLS. The final transfer function is a WA of updated optima, but in contrast to WA, optima are updated by considering residual correlations in the biological data. In practice, taxa abundant in samples with large residuals are most likely to have updated optima (ter Braak and Juggins, 1993).

The main advantage of WA-PLS is that it usually produces models with lower RMSEP and lower bias than WA (ter Braak and Juggins, 1993). There are two reasons for this:

1. All WA-based models suffer from 'edge effect' problems that result in inevitable overestimation of optima at the low end of the gradient of interest and underestimation at the high end of the gradient. As a result there is a bias in the inferred values and in the residuals. WA-PLS implicitly involves a weighted inverse deshrinking regression that pulls the inferred values towards the training set mean. WA-PLS exploits patterns in the residuals to update the transfer function, thereby reducing errors and patterns in the bias.
2. In real life there are often additional environmental predictors that influence the biological assemblages. WA ignores structure resulting from these variables and assumes that environmental variables other than the one of interest have negligible influence. WA-PLS uses this additional structure to improve estimates of the taxa 'optima' in the final transfer function. For optimal performance, the joint distribution of these environmental variables in the past should be the same as in the modern data (ter Braak and Juggins, 1993).

The main disadvantage of WA-PLS compared to WA is that great care is needed in model selection. As more components are added, the WA-PLS model seems to fit the data better as the root mean square error (RMSE) decreases and becomes 0 when the number of components equals the number of samples. RMSE is an 'apparent' statistic of no predictive value as it is based on the training set alone. An independent test set is needed to evaluate different models as the optimal model is the model giving the lowest RMSEP for the test set. In real life there are usually no independent test sets and model evaluation is based on cross-validation to derive approximate estimates of RMSEP. In leave-one-out cross-validation, the WA-PLS modelling procedure for 1, ..., p components, where p is less than n (usually 6–10), is applied n times using a training set of size $(n-1)$. In each of the n models, one sample is left out and the transfer function based on the $(n-1)$ modern samples is applied to the one test sample omitted from the training set, giving a predicted x for that sample. This is subtracted from the known value to give a prediction error for that sample. Thus, in each model, individual samples act in turn as a test set, each of size 1. The prediction errors are accumulated to form a 'leave-one-out' RMSEP. The final WA-PLS model to use in reconstruction is selected on the basis of low RMSEP, small number of 'useful' components (a 'useful' component gives a RMSEP reduction of ≥ 5 per cent of the one-component model) (Birks, 1998), and low mean and maximum bias (ter Braak and Juggins, 1993; Birks, 1995). Sample-specific errors are estimated by cross-validation and Monte Carlo simulation (Birks, 1998).

Like WA, WA-PLS performs surprisingly well and considerably better than direct analogue-matching procedures when none of the fossil assemblages are similar to the modern data (ter Braak *et al.*, 1993; ter Braak, 1995). For very strong extrapolation beyond the modern training set, WA may perform better than WA-PLS. Like WA, WA-PLS is an indicator species approach but where all taxa are used in the transfer function and estimates of the relevant taxon parameters (beta regression coefficients or 'optima') are derived from the modern training set rather than from modern autecological observations.

9.4 EVALUATION AND VALIDATION OF PALAEOENVIRONMENTAL RECONSTRUCTIONS

This has received surprisingly little attention. It is important as all reconstruction procedures will produce results. How reliable are the results?

The most powerful validation is to compare reconstructions, at least for the recent past, against recorded historical records (e.g. Fritz *et al.*, 1994; Lotter, 1998). An alternative approach compares reconstructions with independent palaeoenvironmental data, for example by comparing pollen-based climate reconstructions with plant macrofossil data (Birks and Birks, pp. 000–000 in this volume), pollen-based climate reconstructions with stable-isotope stratigraphy (Hammerlund *et al.*, 2002), chironomid-based climate reconstructions with plant macrofossil data (Brooks and Birks, 2000), etc. Such comparisons are part of the importance of multi-proxy approaches (Lotter, pp. 000–000 in this volume). Without historical validation or independent palaeoenvironmental data, evaluation must be done indirectly using numerical criteria (Birks, 1998).

There are four useful numerical criteria (Birks, 1998):

1. sample-specific RMSEP for individual samples
2. 'goodness of fit' statistics assessed by fitting fossil samples 'passively' onto the ordination axis constrained by the environmental variable being reconstructed for the modern training set and evaluating how well individual fossil samples fit onto this axis in terms of their squared residual distance (Birks *et al.*, 1990)
3. 'analogue' measures for each individual fossil sample in comparison with the training set. A reconstructed environmental variable is likely to be more reliable if the fossil sample has modern analogues within the training set (ter Braak, 1995)
4. the percentages of the total fossil assemblage that consist of taxa (a) that are not represented at all in the training set and (b) that are poorly represented (e.g. < 10 per cent occurrences) in the training set and hence whose transfer function parameters (WA optima, WA-PLS beta-coefficients, etc.) are poorly estimated and have high associated standard errors in cross-validation (Birks, 1998).

Hammarlund *et al.* (2002) and Bigler *et al.* (2002) illustrate these evaluation criteria in multi-proxy studies on Holocene climatic change in northern Sweden.

9.4.1 Computer Software

As transfer functions and reconstruction diagnostics are computer-dependent, the relevant DOS/Windows software is:

Two-way WA	– WACALIB (Line <i>et al.</i> , 1994), CALIBRATE (Juggins and ter Braak, 1997a).
Sample-specific errors in WA	– WACALIB
WA-PLS and leave-one-out cross-validation	– CALIBRATE, WAPLS (Juggins and ter Braak, 1997b).
Sample-specific errors in WA-PLS	– WAPLS
Analogue statistics	– ANALOG (unpublished program by J.M. Line and H.J.B. Birks), MAT (Juggins, 1997).
Reconstruction goodness-of-fit statistics	– CANOCO 4.5 (ter Braak and Šmilauer, 2002).
Other reconstruction diagnostics	– CEDIT (program by Onno van Tongeren supplied with CANOCO 4.0).

For details of availability, see <http://www.campus.ncl.ac.uk/staff/stephen.juggins> (CALIBRATE, WAPLS, MAT) and <http://www.microcomputerpower.com> (CANOCO, CEDIT), or e-mail John Birks (WACALIB, ANALOG) (John.Birks@bot.uib.no).

9.4.2 Limitations

The major limitation is the quality and internal consistency of the modern and fossil data sets. Such sets require a detailed and consistent biological taxonomy and, for the modern data sets, reliable and representative environmental data. The creation of modern training sets with detailed and consistent taxonomy ideally requires that all the biological analyses be done by the same analyst who must be skilled in the relevant taxonomy. Many training sets covering broad geographical areas and ecological gradients have, by necessity, to be constructed from samples analysed by different analysts (e.g. Seppä and Birks, 2001). In such cases, taxonomic workshops, standardization of methodology, quantitative analytical quality control and agreed taxonomic and nomenclatural conventions are essential. Such harmonization between data sets and analysts is time-consuming and unattractive and as a result is underfunded or even bypassed as national, continental and global computer databases rapidly develop. Even less attention is often given to be the quality and representativeness of the modern environmental data used. Particular problems arise in deriving reliable climate data for mountainous areas (e.g. Korhola *et al.*, 2001). Standardized procedures of interpolation, lapse-rate corrections, etc. must be used throughout.

Transfer functions generally work well in the Late Glacial (e.g. Birks *et al.*, 2000) where environmental changes are large and exceed the inherent sample-specific errors of prediction for individual fossil samples (c. 0.8–1.5 °C) (e.g. Brooks and Birks, 2001). In the Holocene these errors are near to the likely range of temperature change and interpretation is correspondingly more difficult (e.g. Korhola *et al.*, 2000). An important area for future research (see below) concerns reducing these errors by adopting a Bayesian approach for environmental reconstruction (Korhola *et al.*, 2002) or by using local training sets (Birks, 1998).

An inherent limitation of all unimodal-based reconstruction methods using WA estimation (WA, WA-PLS) is the 'edge effect' that results in distortions at the ends of the environmental gradient (ter Braak and Juggins, 1993). Although the implicit inverse regression in WA-PLS helps reduce edge effects, it has its own problems by 'pulling' the predicted values towards the mean of the training set, resulting in an inevitable bias with overestimation at low values and underestimation at high values. At present there seems no way to reduce the truncation of taxon responses and hence under- or overestimation of optima, except by using shorter environmental gradients, linear-based methods and local training sets (Birks, 1998).

A further problem arises in interpreting quantitative palaeoenvironmental reconstructions. They are invariably rather 'noisy' resulting, in part at least, from the inherent sample-to-sample variability in the biostratigraphical data used. Non-parametric regressions such as locally weighted regression smoothing (LOWESS) provide useful graphical tools for highlighting 'signal' or major patterns in time-series of reconstructed environmental variables. LOWESS (Cleveland, 1993) models the relationship between a dependent variable (e.g. pollen-inferred July temperature) and an independent variable (e.g. age) when no single functional form such as a linear or quadratic model is appropriate. LOWESS provides a graphical summary that helps assess the relationship and detect major trends within 'noisy' data or reconstructions. LOWESS can also be used to provide a 'consensus' reconstruction based on several reconstructions (e.g. different transfer functions or different proxies) (Birks, 1998). Examples include Bartlein and Whitlock (1993), Lotter *et al.* (1999, 2000) and Fig. 9.3. More sophisticated smoothers such as SiZer (Chaudhuri and Marron, 1999) have considerable potential in palaeoecology because they assess which features seen in a range of smooths are statistically significant (see Korhola *et al.*, 2000).

BJÖRNFJELLTJÖRN
Selected pollen & spore percentages
Anal: Sylvia M. Peglar, 1995

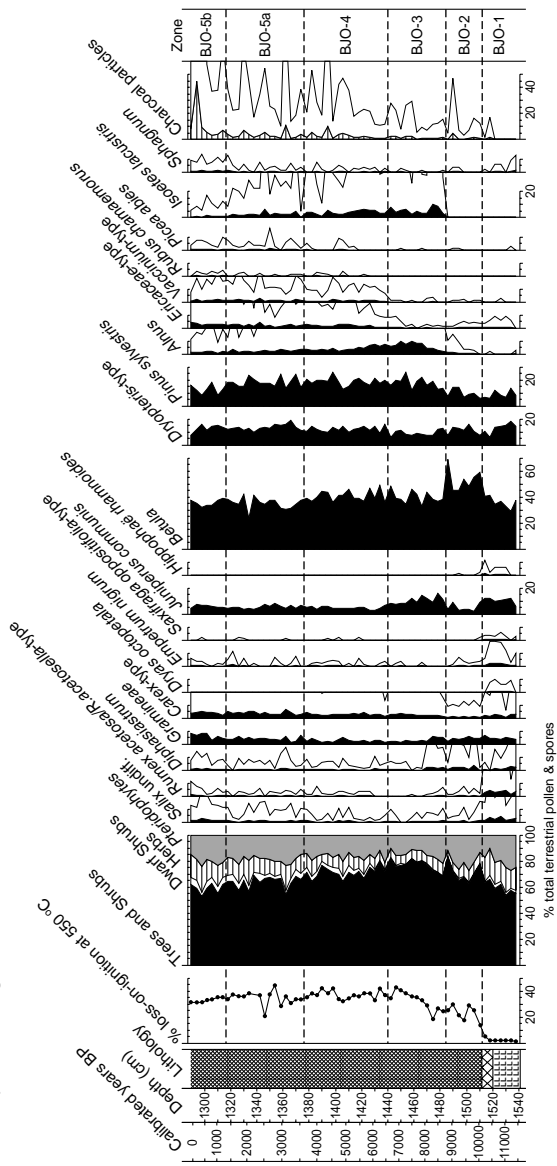


Figure 9.3 Summary pollen diagram from Björnfjelltjörn, north Norway showing major taxa only. The pollen and spore values are expressed as percentages of total land pollen and spores. The unshaded curves are x10 exaggeration. An age-scale based on eight AMS ^{14}C dates and an associated age-depth model for the core is also shown.

9.5 AN APPLICATION

To illustrate using different environmental reconstruction techniques in Holocene palaeoclimatology, WA, WA-PLS and MAT are applied to the Holocene pollen stratigraphy from Björnfejelltjörn, northern Norway (Fig. 9.4). This is a small 3 ha lake at 510 m above sea level near the Norwegian–Swedish border east of Narvik at 68°26'N latitude and 18°04'E longitude. A 2.9 m long core of sediment was obtained from the deepest (12.9 m) part and the core analysed for pollen, spores and plant macrofossils by Sylvia M. Peglar. Eight AMS ^{14}C dates were obtained and an age-depth model developed using a weighted non-parametric regression procedure. The site occurs just above present-day *Betula pubescens* forest limit and lies within the low-alpine region. Vegetation cover is patchy with much bare rock, dwarf-shrub heath dominated by *Betula nana* and heaths (e.g. *Vaccinium myrtillus*, *Arctostaphylos alpinus*) and snow-beds dominated by *Salix herbacea*. Present-day mean July temperature is 10.5 °C.

Holocene mean July temperatures were reconstructed using MAT, WA and WA-PLS and a 191-sample training set of surface-mud samples from throughout Norway and northern Sweden (Seppä and Birks, 2001). The training set covers 7.7–16.4 °C mean July temperature, has 152 pollen and spore taxa, and a RMSEP of 1.07 °C (WA), 1.33 °C (MAT) and 1.03 °C (WA-PLS 3 components).

The reconstructions (Fig. 9.5) are based on MAT, WA, WA-PLS and all 98 terrestrial fossil pollen and spore taxa. For comparability the y -axis is plotted on the same scale throughout and the reconstructions are plotted on a calibrated (cal) age scale (years BP) based on the age-depth model for the core. In addition the WA-PLS reconstruction and associated sample-specific errors are shown on Fig. 9.5. LOWESS smoothers are fitted to each reconstruction to highlight major trends. The MAT suggests little change throughout the Holocene, whereas WA suggests a rise of about 1 °C between 5000 and 9000 cal years BP. WA-PLS suggests larger climatic shifts, with mean July temperatures about 1.5 °C warmer than today between 6000 and 9000 cal years BP and marked changes in the Late Holocene. Sample-specific errors are about 1.0–1.2 °C, but are large relative to the magnitude of change in the reconstructions. Numerical evaluation of the individual reconstructions suggest that they are reliable on statistical criteria, with low residual distances, good analogues, almost all fossil taxa well represented in the training set and consistent sample-specific errors.

A consensus reconstruction (Fig. 9.3) is derived by fitting a LOWESS smoother through all reconstructed values (MAT, WA, WA-PLS and also PLS). This consensus highlights warmer July temperatures than today from about 4500 to 9000 cal years BP. It should be validated using an independent proxy, in this case plant macrofossils that are not used in deriving the pollen-based temperature reconstructions. This is presented by Birks and Birks (pp. 000–000 in this volume).

9.6 CONCLUSIONS AND POSSIBLE FUTURE DEVELOPMENTS

The major conclusion is that quantitative Holocene palaeoenvironmental reconstructions are possible from biostratigraphical data but we are probably near the resolution of current data and methods, with sample-specific errors of 0.8–1.5 °C. The transfer function approach is dependent on modern and fossil data of high taxonomic and analytical quality. The acquisition

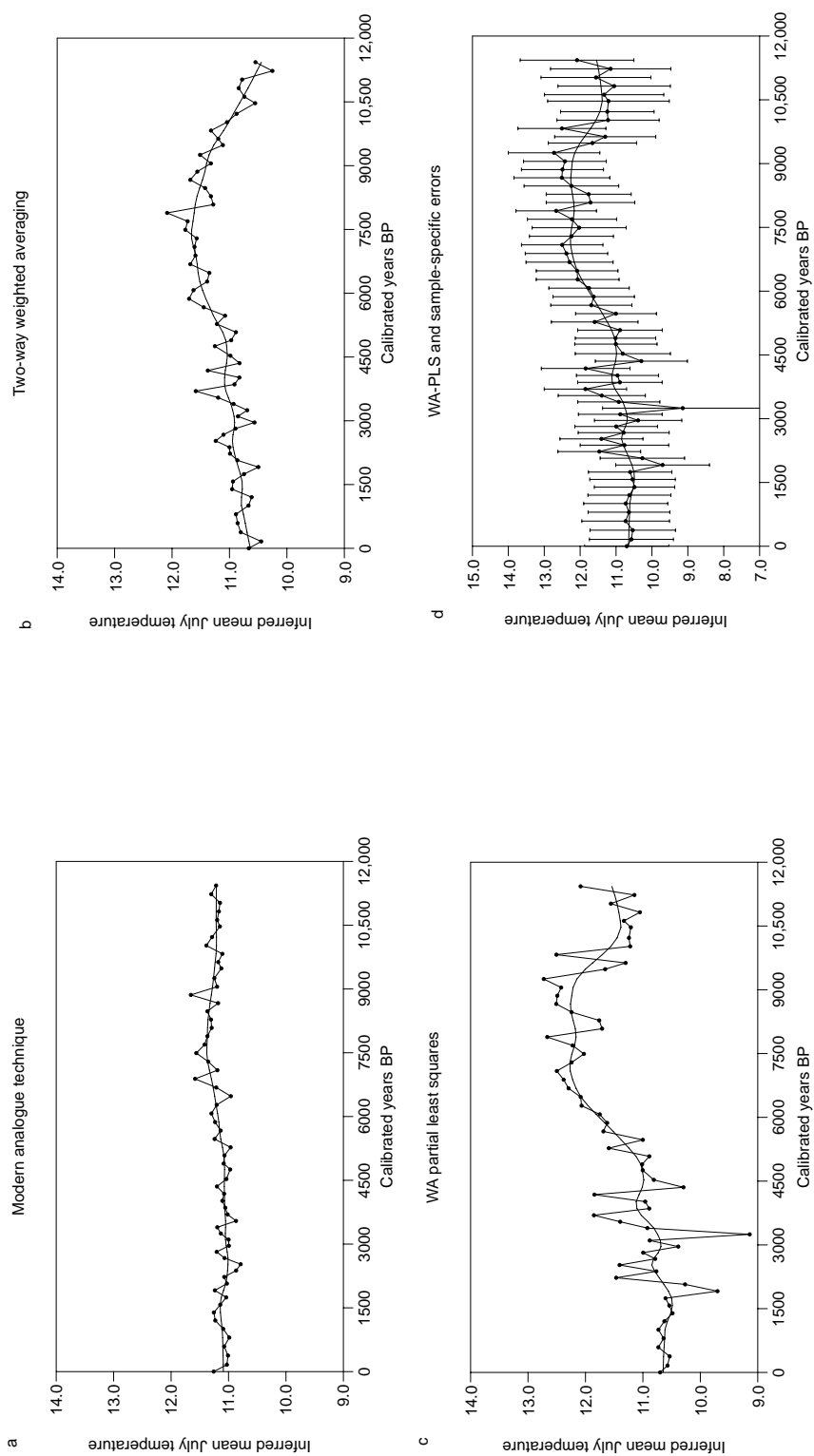


Figure 9.4 Reconstructions of mean July temperature (°C) for Björnfeldtjörn based on modern analogue technique, two-way weighted averaging, and weighted-averaging partial least squares (WA-PLS). The reconstructed values are joined up in chronological order and a LOWESS smoother (span = 0.2, order = 2) has been fitted to highlight the major trends. In the lower right-hand plot, the sample-specific errors of reconstruction for the WA-PLS reconstruction are also shown. All reconstructions are plotted against the age scale (calibrated years BP) of Figure 9.3.

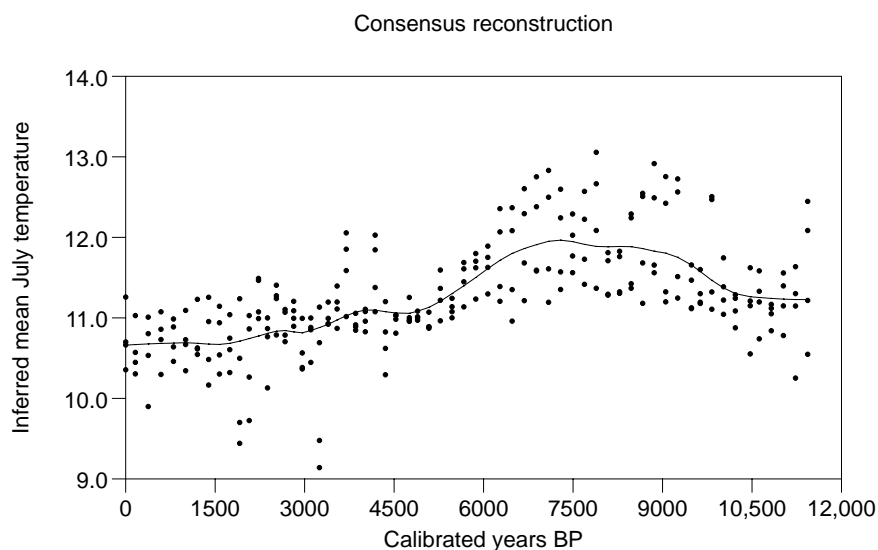


Figure 9.5 Consensus reconstruction for mean July temperature ($^{\circ}\text{C}$) at Björnfielljtjörn based on modern analogue technique, two-way weighted averaging, weighted-averaging partial least squares regression (see Fig. 9.4), and partial least squares regression. The consensus reconstruction is shown as the fitted line and is a LOWESS smoother (span = 0.2, order = 2) fitted through all the reconstructed values.

of such data is time-consuming. A further challenge in Holocene palaeoenvironmental research is to refine transfer function methodology, reduce sample-specific errors, and distinguish 'signal' from 'noise' in reconstructions. Such improvements may come about in three ways: (1) improvements in the quality and reliability of the modern environmental data; (2) improvements in numerical methods for deriving transfer functions reconstructions; and (3) improvements in interpreting palaeoenvironmental reconstruction time-series.

It is difficult to see how to improve the modern climatic data used in training sets given the availability and quality of modern climate data and the complex patterns of climate variation over small areas, especially with complex topography. Geographical information systems and spatially-explicit statistical modelling and interpolation procedures (e.g. Fotheringham *et al.*, 2000) have the potential for improving climatic data for modern transfer functions.

Improvements in numerical methods may come from artificial neural networks and adopting a Bayesian framework. Applications of artificial neural networks in palaeoceanography (Malmgren and Nordlund, 1997; Malmgren *et al.* 2001) and palaeolimnology (Racca *et al.* 2001, 2002) show their potential in Holocene research. The Bayesian approach does not rely on an explicit model of relationships between variables but on the modification of some prior belief about the specific value of a variable on the basis of some additional information (Robertson *et al.*, 1999). In Bayesian terms, this is known as the prior probability. This can be refined with additional information provided by the modern training set, for example, to give the conditional probability. Once the conditional probability density function has been obtained, it can be combined with the prior probability density function to provide a posterior probability density

function using Bayes' theorem (Robertson *et al.*, 1999). Recent studies (Vasko *et al.*, 2000; Toivonen *et al.*, 2001; Korhola *et al.*, 2002) indicate potential advantages of developing transfer functions within a Bayesian framework, although the current computing demands are beyond the computing facilities generally available to palaeoecologists.

Despite considerable advances in transfer function methodology and in developing organism–environment training sets, our abilities to interpret and compare time-series of palaeoenvironmental reconstructions have hardly developed beyond visual comparisons of time-series (Bennett, 2002). There is great scope for applying robust approaches for comparing time-series (e.g. Burnaby, 1953; Malmgren, 1978; Schuenemeyer, 1978; Malmgren *et al.*, 1998). Newly developed techniques for spectral and cross-spectral analysis and of unevenly spaced time-series that are so frequent in palaeoecology (Schulz and Stattegger, 1997; Schulz and Mudelsee, 2002) have the potential for permitting a critical and statistically rigorous interpretations of Holocene palaeoenvironmental time-series. These are major challenges for the future.

Acknowledgements

I am grateful to Sylvia M. Peglar for providing pollen data from Björnffjelltjörn and the modern pollen data set, to Arvid Odland for providing modern climate data, to Hilary Birks, Steve Juggins, Andy Lotter, Heikki Seppä and Cajo ter Braak for helpful discussions about palaeoenvironmental reconstructions, and to Anne Birgit Ruud Hage, Sylvia Peglar and Christopher Birks for help in preparing the manuscript.