4 Calibration

C.J.F. ter Braak

4.1 Introduction

In Chapter 3, we used regression analysis to analyse the way in which species respond to environmental variables. The goal of regression analysis is to express the response of a species as a function of one or more environmental variables. In this chapter, we consider the reverse problem: namely how to express values of an environmental variable as a function of species data. This function is termed the 'transfer function' or 'biotic index' and its construction is termed calibration. The calibration problem differs from the regression problem, because the causal and statistical relations between species and environment are asymmetric.

It might be thought easier to measure environmental variables at a site than to infer their values from the species that occur there. But often it is not. For example, total values over time may be required; repeated measurements are costly, while species automatically integrate environmental conditions over time. This is one of the ideas behind biological evaluation of water quality and bio-monitoring in general. There are also situations where it is impossible to measure environmental variables by direct means, whereas a biological record does exist. An example is the reconstruction of past changes in acidity (pH) in lakes from fossil diatoms from successive strata of the bottom sediment.

An indicator species is ideally a species that always occurs under a unique set of environmental conditions and does not occur elsewhere. Such an ideal indicator species indicates its unique set of environmental conditions without error. Ideal indicator species do not exist, however. Species with narrow ecological amplitudes exist, but such species are not always present in their specific environment and many of them have a low probability of occurrence there, partly because we do not know their specific environmental requirements fully. If such species occur somewhere, they indicate the environmental conditions at that place precisely, but their absence provides hardly any information about the environment. This is a major reason to use the whole community composition at a site for calibration purposes, including species with wider ecological amplitudes. In practice, 'community composition' is restricted to mean species of a particular taxonomic group, e.g. diatoms or vascular plants. Our definition of indicator species is broader than the one used in standard bioassay applications, where individuals of a single species are put on test to determine the amount of some drug or pollutant. Environmental calibration can, however, be considered as a multi-species form of bioassay.

In this chapter, we will introduce three calibration methods, one based on response functions, one on indicator values and one on inverse regression. In the first method (Section 4.2), the response functions can be of any type, whereas in the other two methods particular response curves are assumed, unimodal curves in the one based on indicator values (Section 4.3) and straight lines in the inverse regression method (Section 4.4).

4.2 Maximum likelihood calibration using response functions

4.2.1 Introduction

Maximum likelihood calibration is based on response functions of species against environmental variables. We shall assume that these functions are known, i.e. they have already been estimated from an appropriate and sufficiently large set of data by regression analysis (Chapter 3). (This set of data is termed the training set.) For each set of values of environmental variables, we thus know what the probability is of observing a particular species composition. What we want is to predict the set of values of environmental variables at a particular site from its species composition. When the maximum likelihood principle is used, the prediction is the set of values that would give the maximum probability of observing that particular species composition, if that set of values were the true condition (cf. Subsection 3.3.2). This principle is illustrated in Subsection 4.2.2, together with the concept of a prior distribution and the loss in efficiency when ignoring possible correlations between species. In Subsection 4.2.2, we consider the problem of predicting a nominal environmental variable from presence–absence species data. This type of calibration is also known as discriminant analysis. How to discriminate between classes of a nominal variable by using abundance data will be discussed in the next chapter, in Subsection 5.5.5. In Subsection 4.2.3, the maximum likelihood principle is used to predict values of a quantitative environmental variable, first from presence–absence species data and then from abundance data.

One is commonly interested in a single environmental variable, whereas the species might respond to many more environmental variables. This problem can be solved in maximum likelihood calibration by using response functions of all the important environmental variables; the principles remain the same. But the response functions have first to be estimated from data by regression (Chapter 3), and the size of the training set of data will put a limit on the number of environmental variables that can be taken into account.

4.2.2 Predicting a nominal environmental variable

As an example, suppose we want to estimate the unknown value of soil type from the presence of a particular species. Let us assume that soil type has three classes, clay, peat and sand, and that the probabilities that the species occurs on a field of a given size are 0.1 for clay, 0.2 for peat and 0.4 for sand. If this species is encountered, then the maximum likelihood estimate of the soil type is sand, because sand is the soil type on which the species occurs with the highest probability. If the species is absent, the maximum likelihood estimate is clay, because clay is the soil type where the species is absent with the highest probability. These are the rules of assignment or classification. When the species is present, the proportion of wrong assignments is $(0.1 + 0.2)/(0.1 + 0.2 + 0.4) = 0.43$. If the species is absent, the proportion of wrong assignments is $(0.8 + 0.6)/(0.9 + 0.8 + 0.6) = 0.61$, a small reduction compared to random assignment, so then the assignment procedure is not very effective; note also that the rules defined above never assign to soil type peat.

In these rules, it was implicit that clay, peat and sand occurred equally frequently. This may not be so. If we know beforehand that soil type clay is encountered three times as often as peat or sand, then we could bet on the soil type being clay without any further information. This knowledge about the soil types a priori is termed the 'prior distribution', which is 0.6, 0.2, 0.2 in the example. If we also know that the species is present in a particular field, the probability that its soil type is clay is (apart from a normalizing constant) the product of the prior probability of clay and the probability that the species occurs on clay, that is: $0.6 \times 0.1 = 0.06$, compared to $0.2 \times 0.2 = 0.04$ for peat and $0.2 \times 0.4 = 0.08$ for sand. From these values, we obtain 'posterior probabilities' by dividing these values by their sum, $0.06 + 0.04 + 0.08 = 0.18$ in the example, so that the posterior probabilities for clay, peat and sand are 0.33, 0.22 and 0.44, respectively. The maximum of these probabilities is 0.44 for sand. The extra information that the species is present in the field changes our preference a priori from clay to sand. If the prior distribution is, however, 0.8, 0.1 and 0.1, then the maximum likelihood estimate is always clay, even if the species is present at the field. It is therefore important for the construction of the assignment rule for what frequencies the soil types are expected to be encountered on when the assignment rule will be used. The prior distribution is said to be uniform when the frequencies are equal. This distribution is often assumed when the true distribution is unknown. Many of the controversies in the statistical literature about calibration concern the question whether it is prudent to use the distribution of the fields in the training set as a prior distribution (Brown 1979). Rules based on the maximum likelihood principle have the attractive property that they minimize the number of wrong assignments (misclassifications). As a consequence, each wrong assignment is counted equally. There are, however, situations where one wrong assignment (e.g. assignment to peat instead of to clay) has more serious consequences than another (e.g. assignment to peat instead of to sand). This aspect of costs can be incorporated in the construction of assignment rules (e.g. Lachenbruch 1975).

In the following, we will assume equal costs for wrong assignments and a uniform prior distribution unless explicitly stated otherwise. So environmental conditions will be predicted on the basis of the response function of the species only.

We now extend the example. Apart from the species of that example, Species A, there is a second species, Species B, that only occurs rarely on clay or peat ($p = 0.01$) but often on sand ($p = 0.98$). If a field only contains Species A, then the absence of Species B indicates that its soil type is not likely to be sand; peat is then the most logical bet. Peat is also the maximum likelihood estimate

if the responses of the species are independent; the probabilities of 'Species A present and Species B absent' for the three soil types are $0.1 \times 0.99 = 0.099$, 0.198 and 0.008, respectively, the maximum being for peat. The proportion of wrong assignment (0.35) is less than in the first example with Species A only. In this example (and also in the previous one), the absence of a species thus provides information on the environment.

In this example, an extra assumption was needed to calculate the probability of 'Species A present and Species B absent', namely that the responses of the two species were independent, so that the joint probability could simply be obtained by multiplication of the probability of 'Species A present' and the probability of 'Species B absent'. However the example was constructed in such a way that the best assignment rule would not change, even if the responses of the species were interdependent. In the next example, the assignment rule can be improved considerably if we account for known correlation between the responses of species.

For simplicity, this example includes only two soil types, clay and sand, with equal probabilities of occurrence of Species A ($p = 0.2$) and of Species B ($p = 0.4$). If the responses of Species A and Species B are independent, there is no way of discriminating between clay and sand on the basis of their responses; each assignment rule is wrong for half the cases. But suppose now that these species have preference for a different water-table when on sand, and are indifferent to the water-table when on clay. If both species are encountered in a field, its soil type is not likely to be sand. The probability of both species being present is close to zero on sand, whereas this probability is much larger on clay ($0.2 \times 0.4 = 0.08$). It is therefore possible to improve the assignment rule by using the (negative) correlation between the species. To construct this improved rule, we must know four probabilities:

- the probability of A only
- the probability of B only
- the probability of A and B
- the probability of neither A nor B.

If there are m species, we need to know 2^m probabilities to construct the maximum likelihood assignment rule. All these probabilities must be estimated from the training set, an impossible task if the number of species exceeds 10, even if the training set is huge. Lachenbruch (1975, p. 41-46) described solutions to this problem when the dependence between species is simple. If the dependence between responses is caused by another environmental variable, it is most natural to incorporate this variable explicitly in the response function and to maximize the likelihood for both environmental variables jointly.

4.2.3 *Predicting a quantitative environmental variable*

Presence-absence species data

Assume that the response curve of the probability that a particular species is present is unimodal. Further assume that the environmental variable to be inferred takes the value x_0 for a particular field. If the species is present, the maximum likelihood estimate of x_0 is then the optimum of the curve. At the optimum, the probability of occurrence of the species is clearly maximum. If the species is absent, there are two maximum likelihood estimates, $-\infty$ and $+\infty$.

Suppose now that there are m species that respond to a single quantitative environmental variable x only and suppose that the responses of the species are mutually independent for each fixed value of x. Denote the response curve of the probability of occurrence of the k-th species by $p_k(x)$. The probability that the k-th species is absent also depends on x and equals $1 - p_k(x)$.

The probability of a combination of species is, by their independence, the product of the probabilities of occurrence of the species that are present and the probabilities of absence of the species that are absent. The maximum likelihood estimate of x_0 is, again, the value for which the probability of the observed combination of species is maximum. In principle, we can calculate this probability for any value of x and determine the value of x that gives the highest probability. In practice, we need to write a computer program to do so.

The ratios of probabilities for different values of x, and not the absolute probabilities, are relevant in the estimation because a product of probabilities is calculated for every value of x. For rare species, whose maximum probability of occurrence is small, the ratio of the probabilities of occurrence for two values of x can still be very large. But the probability that a rare species is absent is always close to 1, irrespective of the value of x. The ratio of the probabilities of absence for different values of x is therefore always close to 1. Consequently, absences of rare species cannot influence the maximum likelihood estimate very much and so provide hardly any information on the environment at a site.

Quantitative abundance data

We now consider the estimation of an unknown value of a quantitative environmental variable (x) from a quantitative response (y) of a single species. If the response function is Ey = f(x) and the error is normally distributed, we obtain the maximum likelihood estimate by solving the equation $y = \mathrm{f}(x_0)$ for x_0. In a graph of the response curve, this simply means drawing a horizontal line at the level of the value y and reading off x where this line cuts the response curve. For the straight line (Figure 3.1), this gives the estimate

$$\hat{x}_0 = (y - b_0)/b_1.$$

If the response curve is unimodal, the horizontal line cuts the response curve twice so that we obtain two estimates. This problem has led de Wit et al. (1984) to suggest that an indicator species should have a monotonic relation with the environmental variable of interest. But, if more than one species is used for calibration, the problem generally disappears (Brown 1982).

For later reference (Subsection 5.3.2), we consider the case where each of m species shows a straight-line relation with x, and we want to predict x_0 from the m abundance values at the site. Reading off the graph for each species would give m possibly different estimates of x_0, and we want to combine them. The model for the data can be written as

$$\mathrm{E}y_k = a_k + b_k x \qquad \text{Equation 4.1}$$

where
y_k is the response of species k,
a_k its intercept and
b_k its slope parameter.

By minimizing the sum of squares of differences between the observed and expected responses, we obtain the combined estimate (as Equation 3.6):

$$\hat{x}_0 = \Sigma_{k=1}^{m}(y_k - a_k)b_k / \, \Sigma_{k=1}^{m} b_k^2 \qquad \text{Equation 4.2}$$

This is the maximum likelihood estimate only in the special case that the species are independent and have equal error variances. For the general case see Brown (1982).

4.3 Weighted averaging using indicator values

In this calibration method, the relation between a species and a (semi-) quantitative environmental variable (x) is summarized by a single quantity, the indicator value. Intuitively, the indicator value is the optimum, i.e. the value most preferred by a species. The value of the environmental variable at a site (x_0) is likely to be somewhere near the indicator values of the species that are present at that site. The method of weighted averaging takes it to be the average of these indicator values. If we have recorded abundances of the species, we may take a weighted average with weighting proportional to species' abundance and absent species carrying zero weight. The weighted average of indicator values is thus

$$\hat{x}_0 = (y_1 u_1 + y_2 u_2 + \ldots + y_m u_m) \, / (y_1 + y_2 + \ldots + y_m) \qquad \text{Equation 4.3}$$

where
$y_1, y_2, \ldots, y_m$ are the responses of the species at the site,
$u_1, u_2, \ldots, u_m$ are their indicator values.

For presence–absence data, the average of the indicator values of the species present is also called 'weighted' because absent species implicitly carry zero weight. Note that the method of weighted averaging is also used in Section 3.7 to estimate the indicator value of a species, in particular, by taking a weighted average of values of an environmental variable (Equation 3.28).

The weighted average was proposed as a biotic index for many types of organisms: for vascular plants by Ellenberg (1948) and by Whittaker (1956); for algae by Zelinka & Marvan (1961); and for faunal communities in streams and rivers by Chutter (1972). A typical example is Ellenberg's (1948; 1979) system for predicting soil acidity, reviewed by Böcker et al. (1983). Ellenberg has grouped Central European plants into nine preference groups of soil acidity and assigned the scores 1 to 9 to these groups, the score 1 to the group with species that preferred the most acid conditions and the score 9 to the group with species that preferred the most alkaline conditions. Ellenberg based this grouping on his field observations of the conditions under which particular species occurred and, to a lesser extent, on laboratory tests. The scores are thus the indicator values and are used to derive site scores by weighted averaging. In Ellenberg's system, the indicator values are ordinal and the resulting weighted average is a semiquantitative estimate of soil acidity. Ellenberg (1979), Rogister (1978) and Vevle & Aase (1980) demonstrated a strong relation between the weighted average for acidity based on plant composition and acidity actually measured in the field, thus confirming the empirical predictive value of the weighted average in Ellenberg's system.

From a theoretical viewpoint, it is surprising that the absent species have been disregarded in the weighted average. Apparently it is supposed that absent species do not provide information on the environment of a site (cf. Subsection 4.2.3). Further, each species is regarded as an equally good indicator in weighted averaging, whereas it is intuitively reasonable to give species with a narrow ecological amplitude more weight than species with a broader ecological amplitude. Ellenberg (1979) circumvented this problem by disregarding indifferent species; they were not assigned an indicator value. Zelinka & Marvan (1961) solved this problem in a heuristic way by assigning species not only an indicator value but also an indicator weight. Finally, because the indicator values are ordinal, calculating averages is a dangerous arithmetic operation; ordinal scale values are rather arbitrary, so they could be transformed monotonically without change of meaning. However the order of weighted averages calculated for different sites can be scrambled by such a transformation.

Ter Braak & Barendregt (1986) provided a theoretical justification of using the weighted average (Equation 4.3). For presence–absence data, the weighted average of indicator values is about as efficient as the maximum likelihood estimate of x_0 if the response curves of the species are Gaussian logit curves (Equation 3.17) with equal tolerances and the species presences are independent and if, in addition:

- either the maximum probability of occurrence is very small for any species so that absent species provide no information on the environment (Subsection 4.2.3)
- or as illustrated in Figure 4.1, the indicator values (optima) are homogeneously distributed over a large interval around x_0
- and the maxima of the response curves of species are equal.

If the condition of equal tolerances does not hold true, we must take a tolerance-weighted version of the weighted average

$$\hat{x}_0 = (\Sigma_{k=1}^{m} y_k \, u_k / t_k^2) / (\Sigma_{k=1}^{m} y_k / t_k^2) \qquad \text{Equation 4.4}$$

to retain high efficiency. Here, t_k is the tolerance of species k (Equation 3.17).

For quantitative abundance data, the method of weighted averaging can be justified analogously (ter Braak & Barendregt 1986). If the abundances follow a Poisson distribution and the response curves are Gaussian curves (Equation 3.8) with homogeneously distributed optima, equal tolerances and maxima, the weighted average again approximates the maximum likelihood estimate. This result may help to decide whether it is prudent to transform to presence–absence before the weighted average is calculated.

The conditions (homogeneously distributed optima, equal tolerances and maxima) together make a species packing model (Figure 4.1). This is an ecological model based on the idea that species evolve to occupy maximally separate niches with respect to a limiting resource. Christiansen & Fenchel (1977, Chapter 3) provide a lucid introduction here. This idea applies also to the occurrence of competing species along environmental variables (Whittaker et al. 1973). Response curves should therefore have minimum overlap.

Despite its theoretical basis, the species packing model is not likely to hold in real life. Nevertheless, the derivation of the weighted average provided above indicates the kind of situation in which the weighted average performs reasonably well. Species may not really be distributed according to the species packing model, but neither are they tightly clumped along environmental gradients; there is usually a fairly even turnover of species along gradients. In addition, Equation 4.4 shows how one can incorporate information on ecological amplitudes in the weighted average.

In lists of indicator values, the values are often expressed on an ordinal scale. For weighted averaging to be useful, the scale values (and, hence, the indicator values) must be chosen such that most species show fairly symmetric response curves. If this can be achieved, the weighted average is an informative semiquantitative biotic index. The method of weighted averaging of indicator values is

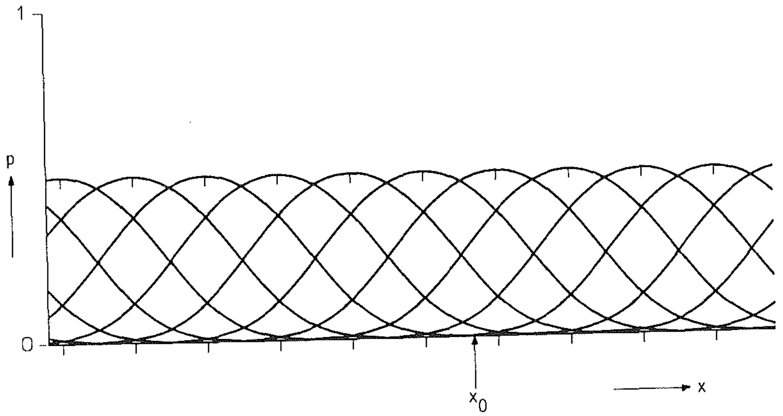

Figure 4.1 Species packing model: Gaussian logit curves of the probability (p) that a species occurs at a site, against environmental variable x. The curves shown have equispaced optima (spacing = 1), equal tolerances (t = 1) and equal maximum probabilities of occurrence ($p_{\max}$ = 0.5). x_0 is the value of x at a particular site.

also attractive to reveal a possible structure in data tables such as Table 0.1 of this book. We simply rearrange the species in order of their indicator value for a particular environmental variable and the sites in order of their weighted average, as in Section 3.7.

4.4 Inverse regression

In Subsection 4.2.3, we discussed a calibration method for when abundance values of a species show a linear relation with the environmental variable of interest. An attractive alternative method is then inverse regression. In inverse regression, the training set is not used to construct response curves by regressing the responses of the species on the environmental variable; instead the environmental variable is taken as the response variable and the responses of the species as the explanatory variable. The regression equation so constructed is then directly the transfer function that is used for prediction. This method has attractive properties if the prior distribution of the environmental variable equals the distribution in the training set (Brown 1979).

The method of inverse regression can easily be extended to prediction on the basis of the responses of more than one species. Each species then makes an explanatory variable, so that the inverse regression is a multiple (least-squares) regression of the environmental variable on the response variables of the species. Predictions are again derived directly from the multiple regression equation so obtained. This method is most efficient if the relation between each of the species and the environmental variable is a straight line with a normal distribution of error (Equation 4.1) and if the environmental variable too has a normal distribution (Brown 1982).

However species do not in general have monotonic relations with environmental variables. For example, response surfaces of pollen types with respect to summer temperature and annual precipitation over large geographic regions are strongly non-linear (Bartlein et al. 1986). Inverse regression could not therefore be used to build one generally applicable transfer function to reconstruct past climates from pollen data. But response curves could be made about linear by limiting the geographic area and transforming the pollen data (Howe & Webb 1983). Therefore Bartlein & Webb (1985) subdivided a large geographic area into regions and, for the actual climatic reconstruction, chose among the transfer functions obtained separately for different regions by using an analogue method (a method to decide to which training set of modern pollen data (i.e. to which region) a fossil pollen sample is most similar). Inverse regression was thus just one step in the whole calibration procedure. A simpler procedure would be to fit non-linear response functions first, as described by Bartlein et al. (1986), and to use these to reconstruct past climates by use of the maximum likelihood principle (Section 4.2).

4.5 Bibliographic notes

The history of the method of weighted averaging has been sketched in Section 4.3. Other biotic indices are listed in Sheenan (1984). Battarbee (1984) reviews various biotic indices for pH reconstruction from diatoms, including one based on inverse regression (see also Davis & Anderson 1985).

Much of the statistical literature on calibration is devoted to the prediction of a single quantitative variable on the basis of a single quantitative response variable, assuming a straight-line relation and a normal distribution of error. Brown (1979) compared the method of inverse regression with the Classical approach by first fitting response functions (Subsection 4.2.3). Calibration with polynomial response functions is treated, for instance, by Scheffé (1973), Schwartz (1977) and Brown (1982). Williams (1959, Chapter 9), Brown (1979), Brown (1982), and Naes & Martens (1984) discuss linear multivariate calibration, the prediction of one or more quantitative variables from more than one quantitative response variable, assuming a linear model.

Discrimination (calibration of a nominal explanatory variable) is treated by Lachenbruch (1975) in a general statistical context, by Titterington et al. (1981) in a medical context and by Kanal (1974) in electrical engineering.

4.6 Exercises

Exercise 4.1 Weighted averaging and maximum likelihood calibration with Gaussian logit curves

With data from Kruijne et al. (1967) on the occurrence of plant species and soil acidity (pH) in meadow fields, ter Braak & Looman (1986) fitted a Gaussian logit curve with respect to pH for each of the species. The curves of seven of the species are shown in Figure 4.2. Their parameters are:

Species name	Code	Optimum	Tolerance	Maximum
Agrostis canina	AC	3.4	1.1	0.84
Stellaria graminea	SG	5.7	0.4	0.38
Alopecurus geniculatus	AG	5.8	0.6	0.58
Plantago major	PM	6.2	0.7	0.34
Bellis perennis	BP	6.4	0.5	0.89
Hordeum secalinum	HS	7.1	0.7	0.57
Glechoma hederacea	GH	8.1	1.5	0.55

Although the parameters were estimated from only 100 fields, we treat them in this exercise as the true parameters. For three meadow fields with a unknown soil acidity, we want to predict the soil acidity from the presences and absences of these seven species. The species that are present are in Field 1 AC, SG and BP, in Field 2 AG and BP, and in Field 3 HS and BP (species not mentioned are absent). Predict the pH of each of these fields:

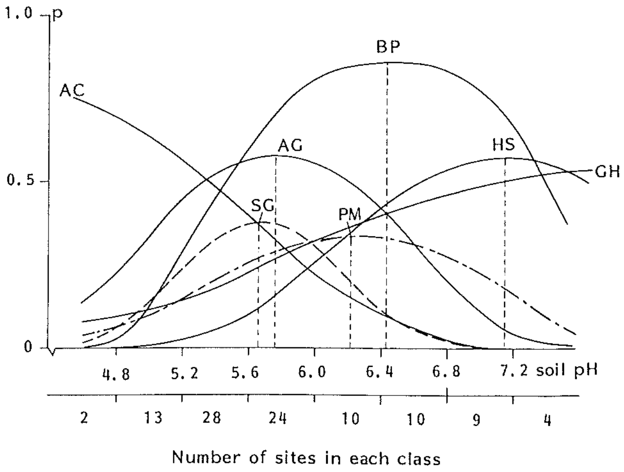

Figure 4.2 Probability of occurrence of seven contrasting species in relation to soil acidity (pH) in meadows, as fitted by logit regression. The curves can be identified by the code near their optimum indicated by dotted lines. The species arranged in order of their optima are: *Agrostis canina* (AC); *Stellaria graminea* (SG); *Alopecurus geniculatus* (AG); *Plantago major* (PM); *Bellis perennis* (BP); *Hordeum secalinum* (HS); *Glechoma hederacea* (GH). Nomenclature follows Heukels–van der Meijden (1983).

Exercise 4.1.1 By the method of weighted averaging using the optima as indicator values.

Exercise 4.1.2 By the tolerance-weighted version of the method of weighted averaging (Equation 4.4).

Exercise 4.1.3 By the method of maximum likelihood. Hint: calculate the likelihood for a limited number of pH values, for example, pH = 5.0, 5.5, 6.0, 6.5, 7.0, 7.5 and next for the most likely value of these plus and minus 0.1. Use Equation 3.17 of Chapter 3 to calculate probabilities of occurrence. In that equation: c = maximum/(1 – maximum).

Exercise 4.2 Calibration using a straight line

Predict, by using the results of Exercise 3.1, the relative sulphate concentration of a moorland pool in which *Frustulia rhomboides* var. *saxonica* occurs with 70 frustules.

Exercise 4.3 Calibration using a Gaussian response curve

Predict, by using the results of Exercise 3.2, the February sea-surface temperatures of two samples in which the abundances of *Spongotrochus glacialis* are 20% and 60%, respectively.

4.7 Solutions to exercises

Exercise 4.1 Weighted averaging and maximum likelihood calibration with Gaussian logit curves

Exercise 4.1.1 The weighted average (Equation 4.3) is for Field 1
$\hat{x}_0 = (1 \times 3.4 + 1 \times 5.7 + 0 \times 5.8 + 0 \times 6.2 + 1 \times 6.4 + 0 \times 7.1 + 0 \times 8.1)/(1 + 1 + 0 + 0 + 1 + 0 + 0) = 15.5/3 = 5.17$.
The prediction is thus pH 5.17. Analogously, the weighted average for Field 2 is 6.10 and for Field 3 is 6.75.

Exercise 4.1.2 The tolerance weighted version of the weighted average (Equation 4.4) gives for Field 1
$\hat{x}_0 = (1 \times 3.4/1.1^2 + 1 \times 5.7/0.4^2 + 0 \times 5.8/0.6^2 + ... + 0 \times 8.1/1.5^2)/(1/1.1^2 + 1/0.4^2 + ... + 0/1.5^2) = 64.03/11.08 = 5.78$. For Field 2 we obtain 6.15 and for Field 3 we obtain 6.64.

Exercise 4.1.3 With Equation 3.17, we obtain the probability of occurrence (p_k) at pH 5.0, which is for AC 0.646, for SG 0.117, for AG 0.362, for PM 0.106, for BP 0.138, for HS 0.015 and for GH 0.126. The probability that the *k*-th species is absent is $1 - p_k$. For pH 5.0, the likelihood of the species combination of Field 1 (AC, SG and BP present) is therefore $0.646 \times 0.117 \times (1 - 0.362) \times (1 - 0.106) \times 0.138 = (1 - 0.015) \times (1 - 0.126) = 0.0051$.
For pH 5.5, 6.0, 6.5, 7.0 and 7.5, we obtain likelihoods of 0.0244, 0.0094, 0.0008, 0.0000, 0.0000, respectively. The maximum of these likelihoods is 0.0244, at pH 5.5. The likelihoods at pH 5.4 and 5.6 are slightly lower and, within the precision of 0.1, 5.5 is the maximum likelihood prediction of the pH of Field 1.

For Field 2, the likelihood at pH 5.0 becomes 0.0121; the maximum (0.083) occurs at pH 6.0. Slightly lower likelihoods are obtained for pH 5.9 and 6.1. The maximum likelihood prediction is thus 6.0.

For Field 3 the likelihood at pH 5.0 becomes 0.0003; the maximum of the six likelihoods occurs at pH 7.0. pH 7.1 gives a slightly higher likelihood, whereas for pH 7.2 the likelihood decreases again. The maximum likelihood prediction is thus 7.1.

Exercise 4.2 Calibration using a straight line

In Exercise 3.1, the regression equation E $\log_e$ (*Frustulia* count + 1) = 5.848 – 5.96 S_{rel} was obtained. In the pool under study, the count is 70, so that $y = \log_e (70 + 1) = 4.263$. Replacing the left side of the regression equation by 4.263, we obtain $S_{rel} = (5.848 - 4.263)/5.96 = 0.27$.

For the sample with 20% *S. glacialis*, we have to solve the quadratic equation -0.00894 temp^2 + 0.2497 temp + 2.119 = log$_e$ (20) = 2.996. There are two solutions, temp = 4.1 °C and 23.8 °C. The temperatures on which the regression equation is based lies between 0.8 and 21.6 °C. If this range is relevant prior information, the prediction of 23.8 °C can be discarded and the remaining prediction is 4.1 °C.

For the sample with 60% *S.glacialis*, the quadratic equation for temperature has no solution. This is not surprising, because the maximum of the Gaussian curve was 48%, which was obtained at 14 °C. The most likely temperature is therefore 14 °C.

5 Ordination

C.J.F. ter Braak

5.1 Introduction

5.1.1 Aim and usage

Ordination is the collective term for multivariate techniques that arrange sites along axes on the basis of data on species composition. The term ordination was introduced by Goodall (1954) and, in this sense, stems from the German 'Ordnung', which was used by Ramensky (1930) to describe this approach.

The result of ordination in two dimensions (two axes) is a diagram in which sites are represented by points in two-dimensional space. The aim of ordination is to arrange the points such that points that are close together correspond to sites that are similar in species composition, and points that are far apart correspond to sites that are dissimilar in species composition. The diagram is a graphical summary of data, as in Figure 5.1, which shows three groups of similar sites. Ordination includes what psychologists and statisticians refer to as multidimensional scaling, component analysis, factor analysis and latent-structure analysis.

Figure 5.1 also shows how ordination is used in ecological research. Ecosystems are complex: they consist of many interacting biotic and abiotic components. The way in which abiotic environmental variables influence biotic composition is often explored in the following way. First, one samples a set of sites and records which species occur there and in what quantity (abundance). Since the number of species is usually large, one then uses ordination to summarize and arrange the data in an ordination diagram, which is then interpreted in the light of whatever is known about the environment at the sites. If explicit environmental data are lacking, this interpretation is done in an informal way; if environmental data have been collected, in a formal way (Figure 5.1). This two-step approach is indirect gradient analysis in the sense used by Whittaker (1967). By contrast, direct gradient analysis is impossible without explicit environmental data. In direct gradient analysis, one is interested from the beginning in particular environmental variables, i.e. either in their influence on the species as in regression analysis (Chapter 3) or in their values at particular sites as in calibration (Chapter 4).

Indirect gradient analysis has the following advantages over direct gradient analysis. Firstly, species compositions are easy to determine, because species are usually clearly distinguishable entities. By contrast, environmental conditions are difficult to characterize exhaustively. There are many environmental variables and even more ways of measuring them, and one is often uncertain of which variables the species react to. Species composition may therefore be a more informative