Weighted Averaging Partial Least Squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration

Cajo J.F. Ter Braak[1,2], Steve Juggins[3,4], H. J. B. Birks[3,4] and Hilko Van der Voet[1,2]

[1] *Agricultural Mathematics Group-DLO, Box 100, 6700 AC Wageningen, the Netherlands;*
[2] *DLO-Institute for Forestry and Nature Research, Box 23, 6700 AA Wageningen, the Netherlands;*
[3] *Environmental Change Research Centre, Department of Geography, University College London, 26 Bedford Way, London WC1H 0AP, UK;*
[4] *Botanical Institute, University of Bergen, Allégatan 41, N-5007 Bergen, Norway*

Citation: ter Braak, C. J. F., Juggins, S., Birks, H. J. B. and van der Voet, H. (1993). Weighted averaging partial least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. Chapter 25 in: *Multivariate Environmental Statistics,* G. P. Patil, and C. R. Rao (eds). Amsterdam: Elsevier Science Publishers B.V. (North-Holland), pp. 525-560.

ABSTRACT

A multivariate calibration method is proposed that relates to correspondence analysis as partial least squares (PLS) relates to principal components analysis. The new method extends and improves the weighted averaging (WA) method for inferring values of environmental variables from biological species compositions, hence the name weighted averaging partial least squares (WA-PLS). WA-PLS is designed to cope with the special features of ecological data, namely, large numbers of species, many zero-abundance values and the non-linear, often unimodal response of species to environmental variables. In simulations of such data in which the species response is governed by two environmental variables, one of which is the variable of interest and the other a nuisance variable, the standard error of prediction (SEP) was reduced by a factor of *ca.* 0.5 by using WA-PLS instead of WA. The length of gradient of the data (between 2 and 8 SD-units) had little influence on the relative performance as had the nuisance variable, except for a short calibration gradient (2 SD). Further comparisons were made with PLS and a maximum likelihood method (MLM) based on the Gaussian response model for compositional data. WA-PLS outperformed PLS and MLM in all cases, the difference with MLM being small if the nuisance variable was unimportant. WA-PLS also compared favourably on a real data example from Imbrie & Kipp (1971) in which sea-surface temperature is reconstructed from fossil foraminiferal composition.

KEYWORDS: Biomonitoring; Canonical Correspondence Analysis; Climate reconstruction; Compositional data; Gaussian response model; Multinomial logit model; No-modern-analogue problem; RC-Association model; Transfer function

1 **INTRODUCTION**

When direct measurement of environmental variables is impossible or too expensive, we need to resort to indirect methods. Species assemblage data may contain the relevant information. Each biological species requires particular environmental conditions for regeneration, establishment and growth. It should therefore be possible to infer the environmental conditions at a site from the species that occur there. This idea is used in the biological evaluation of water quality (Whitton and Rott 1991) and biomonitoring in general (Spellerberg 1991). Fossil species assemblages similarly can provide a record of the palaeoenvironment. Such records are valuable for a historical perspective of current environmental problems such as acid rain (Battarbee 1984) and global warming (Fritz *et al.* 1991, Walker *et al.* 1991a,b) and for testing the accuracy of simulation models such as the general circulation models (COHMAP 1988, Sloan and Barron 1992). The ecological assumptions required for obtaining sound quantitative, palaeoenvironmental reconstructions are listed by Imbrie and Kipp (1971), Birks *et al.* (1990a) and Sloan and Barron (1992). We focus here on the statistical problems.

Inferring environmental variables from species assemblage data is a difficult multivariate calibration problem. It requires a representative training set consisting of data on species assemblages and associated environmental measurements. Species assemblage data have several features that make them special in a statistical sense (Jongman *et al.* 1987):

I)The number of species is often large (10-300) which leads to multicollinearity.

II)The data are either binary (presence/absence) or, if they are quantitative, they still contain many zero values for sites at which the species is absent. Measures of abundance, like counts, density or biomass, are highly variable and always show a skew distribution. Often the data are compositional; the site total is uninformative, if the total count per site is determined by the sampling design.

III)Relationships between species and environmental variables are generally nonlinear. Because of Shelford's law of tolerance (Odum 1971) and niche-space partitioning (Whittaker, Levin and Root 1973), species abundance or probability of occurrence is often a unimodal function of the environmental variables.

Proposed solutions start with Imbrie and Kipp (1971) who reconstructed past sea-surface temperature from assemblages of foraminifera using a form of principal components regression. This can be viewed as a variant of inverse regression (Brown 1982; Bartlein and Webb 1985) that takes account of the multicollinearity problem, if only a few components are utilized in the final regression. Roux (1979) improved the method by replacing principal components analysis by correspondence analysis so as to allow for the unimodality of the species-environment relationship (Ter Braak 1985) and for the compositional nature of the data. Gasse and Tekaia (1983) were concerned with the fact that some information on the species-environment relationship may be lost by retaining only the first few components. This is a potential problem in all two-step methods, known as indirect gradient analysis methods (Ter Braak 1986). Gasse and Tekaia (1983) therefore developed a direct gradient analysis method. In their method the environmental variable was divided into classes. Ter Braak (1986) recognized their method as a special case of canonical correspondence analysis in which there is a single nominal predictor.

But, in canonical correspondence analysis (that is, correspondence analysis constrained by a linear combination of predictor variables, usually environmental variables) predictors can also be quantitative, so that the division into classes is unnecessary. With a single quantitative predictor, canonical correspondence analysis reduces to the weighted averaging method (Ter Braak 1986, 1987: pp. 74-75; Ter Braak and van Dam 1989). This method can be considered as an approximation to the statistically sound, but computationally intensive method of maximum likelihood calibration using Gaussian response curves (Ter Braak 1985; Ter Braak and Barendregt 1986; Birks *et al.* 1990a). Because of its ecological plausibility, its simplicity and its empirical predictive power, the weighted averaging method (WA) has gained popularity, particularly in palaeolimnology (*e.g.* Oksanen *et al.* 1988, Birks *et al.* 1990b, Cumming, Smol and Birks 1991; Dixit, Dixit and Smol 1991; Dixit et al. 1992; Walker *et al.* 1991; Fritz *et al.* 1991; Hall and Smol 1992).

Multivariate calibration methods in chemometrics also started with principal component regression (PCR), but were developed further in another direction (Martens and Naes 1989). Here, PCR led to partial least squares regression (PLS) (Wold *et al.* 1984). Both PCR and PLS use mutually orthogonal components. In PCR, the components are selected to have maximum variance, whereas in PLS they are chosen to maximize the covariance with the regressand (Stone and Brook 1990). Consequently, PLS often needs fewer components and gives lower prediction error than PCR. In this paper, we consider the case where a single variable is to be calibrated. We therefore focus on univariate PLS regression, also known as PLS1. The multivariate version (PLS2) is briefly discussed at the end.

In this paper we develop a PLS regression method from correspondence analysis just as PLS regression can be derived from principal component analysis. We show that, if only the first component is retained, the new method reduces to the weighted averaging method. In the new method, further components utilize the residual structure in the data to improve the fit. Because each component is extracted by weighted averaging, we name the method weighted averaging partial least squares regression (WA-PLS). As in PLS, the number of useful components is estimated on the basis of the standard error of prediction (SEP) estimated by the leave-one-out method.

To evaluate the performance of WA-PLS, we carried out a simulation study and applied the method to the data from Imbrie and Kipp (1971). In the simulations we study the effect of three factors.

I)*Length of gradient.* Length of gradient is a property of an environmental variable in a particular ecological community. If relationships are unimodal, the species' abundance curves rise and fall over shorter or longer ranges, depending on their niche widths. The length of gradient can be defined as the range of the environmental variable divided by the average range of the species. In the simulations, gradient length is varied by varying the species ranges while keeping the environmental range constant. So, narrow niches lead to long gradients. If the species range is larger than the environmental range, unimodal curves may appear sigmoidal or even linear. Is it then advantageous to use PLS instead of WA-PLS?

II)*Influence of a nuisance variable.* The environmental variable we wish to calibrate should be an ecologically important variable in the system, but it will rarely be the only such variable. How do the methods behave if environmental variables other than the one of interest influence the species composition? In the simulations we introduce one other variable and vary the influence of this nuisance variable. The nuisance variable introduces additional, structured noise in the data which, we hope, WA-PLS can exploit.

III)*Number of species.* Weighted averaging methods appear to perform better with species-rich than with species-poor data.

We restrict attention to compositional data, because many of the data for which WA-PLS is designed are expressed as percentages. The simulations were carried out using Minchin's (1987) COMmunity PAttern Simulator (COMPAS). We compare WA-PLS to WA and PLS and to a maximum likelihood calibration method based on the Gaussian response model for compositional data (Ihm and van Groenewoud 1984). The latter is a 'classical' approach to calibration in constrast to the inverse approaches of PLS and WA-PLS. It is a useful standard for comparison because it is the simplest unimodal model. The model is closely related to the RC-association models of Goodman (1986, 1991). With the simulation study we attempt to delimit the domain where WA-PLS can usefully be applied and also to explore, what is termed in palaeo-reconstruction work, the no-modern-analogue problem by simulating sites at environmental conditions that are not represented in the training set.

3 **THEORY**

4.1 **NOTATION**

Let x denote the environmental variable to be calibrated on the basis of a training data set consisting of the environmental vector $\mathbf{x} = (x_i)$, with x_i the value of the environmental variable in site i, and the $n \times m$ matrix $\mathbf{Y} = (y_{ik})$ with y_{ik} the abundance of species k in site i ($y_{ik} \geq 0$) (i = 1...n sites and k = 1...m species). [Note that the use of the symbols x and y is interchanged compared to a part of the literature on the inverse regression approach to calibration, especially PLS.] The estimated or inferred value of x in site i is denoted by $_i$. Let $\mathbf{R} = \text{diag}(y_{1+},...,y_{n+})$ and $\mathbf{K} = \text{diag}(y_{+1},...,y_{+m})$ with y_{i+} and y_{+k} the total abundance in site i and per species k, respectively.

4.3 **WA-PLS: DEFINITION**

In this section we define WA-PLS by the following guideline: "WA-PLS relates to correspondence analysis (CA) as PLS regression (PLS) relates to principal component analysis (PCA)". Let us thus first examine how PLS relates to principal component analysis (PCA) (Stone and Brook 1990). PCA searches for a normed weight vector for the species, **b** say, such that the linear combination $\mathbf{t} = \mathbf{Yb}$ has maximum norm or, in the most common variant of PCA in which the columns of **Y** are centred, maximum variance. This linear combination does not involve the variable to be predicted, **x**, and thus does not need to have any predictive power. In contrast, PLS selects the first component by maximizing the inner product between the linear combination and **x**. The inner product is precisely the covariance in the most common variant in which both **x** and the columns of **Y** are centred beforehand; we treat the centring operation as a particular form of preprocessing and, for greater generality, formulate the techniques without it. As formulae, PCA searches for a weight vector **b** that maximizes

PCA: $\mathbf{t'}\mathbf{t} / \mathbf{b'}\mathbf{b}$, where $\mathbf{t} = \mathbf{Y}\,\mathbf{b}$, (1)

whereas the PLS searches for a weight vector **b** that maximizes

PLS: $\mathbf{t'}\mathbf{x} / \mathbf{b'}\mathbf{b}$, where $\mathbf{t} = \mathbf{Y}\,\mathbf{b}$. (2)

Let us now examine how PCA and CA are related. Whereas PCA searches for optimal linear combinations, CA searches for an optimal weighted average [recall the alternative name of CA, reciprocal averaging (Hill 1973b, 1974, 1982)]. CA searches for a centred and normed weight vector for the species, **u** say, such that the vector of weighted averages, $\mathbf{t} = \mathbf{R}^{-1}\,\mathbf{Y}\,\mathbf{u}$, has maximum variance. Again, this weighted average does not involve the variable to be predicted, **x**, and thus does not need to have any predictive power. According to our guideline for defining WA-PLS, WA-PLS should now select the first component by maximizing the covariance between the vector of weighted averages and **x**. In the formulae, the logic of CA requires the use of weighted norms (**R** and **K** for sites and species, respectively; Greenacre 1984). Thus, CA searches for a centred weight vector **u** ($\mathbf{1}_m'\mathbf{K}\mathbf{u} = 0$) that maximizes

CA: $\mathbf{t'}\mathbf{R}\mathbf{t} / \mathbf{u'}\mathbf{K}\,\mathbf{u}$, where $\mathbf{t} = \mathbf{R}^{-1}\mathbf{Y}\mathbf{u}$, (3)

whereas WA-PLS searches for a centred weighted vector **u** that maximizes

WA-PLS: $\mathbf{t'}\mathbf{R}\mathbf{x} / \mathbf{u'}\mathbf{K}\mathbf{u}$, where $\mathbf{t} = \mathbf{R}^{-1}\mathbf{Y}\mathbf{u}$. (4)

This concludes the definition of the first component of the techniques.

The second, third and later components are chosen to maximize the same criterion as the first component with the additional restriction of orthogonality to earlier components. In CA and WA-PLS, the components are required to be **R**-orthogonal ($\mathbf{t}_i'\mathbf{R}\mathbf{t}_j = 0$ for all $i \neq j$).

The choice of the number of components to use is an essential ingredient of both PLS and WA-PLS. The optimal number minimizes the prediction error of x and is estimated by cross-validation (Wold *et al.* 1984; Stone and Brook 1990).

4.5 **WA-PLS: SOLUTION OF THE MAXIMIZATION PROBLEM**

The maximization problem posed by WA-PLS can be solved as follows. Recall that the solution of CA can be obtained from a general computer program for PCA by pre- and postprocessing of the data (Greenacre 1984). Similarly, the solution of WA-PLS can be obtained from a PLS program by preprocessing not only **Y** (as in CA) but also **x** as follows,

$$\mathbf{Y}^* = \mathbf{R}^{-½}\,\mathbf{Y}\,\mathbf{K}^{-½} \text{ and } \mathbf{x}^* = \mathbf{R}^{½}\,\mathbf{x} \tag{5}$$

and by postprocessing the results of a PLS with $\mathbf{Y}^*$ and $\mathbf{x}^*$ by

$$\mathbf{u} = \mathbf{K}^{-½}\,\mathbf{b}. \tag{6}$$

Thus, no special computer program is required for WA-PLS. When WA-PLS is carried out through a PLS-algorithm, **x** should be **R**-centred on input and the PLS-algorithm should neither centre nor standardize the input variables. The proof is given in the Appendix.

Having defined WA-PLS and an algorithm to carry it out by computer, the questions remain whether it is an improvement and what the final predictor looks like. The first question is answered later on by analyzing simulated and real data. To answer the second question, an explicit algorithm for WA-PLS is given.

Maximizing (4) with the Lagrange multiplier method leads to the weight vector

$$\mathbf{u}_1 = \mathbf{K}^{-1}\mathbf{Y}'\mathbf{x} \qquad (7)$$

and, thus, to the first component of WA-PLS

$$\mathbf{t}_1 = \mathbf{R}^{-1}\mathbf{Y}\mathbf{u}_1. \qquad (8)$$

The fitted values based on the first component are

$$\hat{x}_1 = \alpha_0 + \alpha_1\,\mathbf{t}_1 \qquad (9)$$

obtained by a regression in the **R**-norm of **x** on to $\mathbf{t}_1$, *i.e.* by using weights y_{i+}. In (7) we neglected the centring requirement on $\mathbf{u}_1$ and thus on $\mathbf{t}_1$; but, the centring is immaterial due to the intercept in (9).

The second weight vector $\mathbf{u}_2$ maximizes (4) subject to $\mathbf{t}_2'\mathbf{R}\mathbf{t}_1 = 0$. By the Lagrange multiplier method we obtain

$$\mathbf{u}_2 = \mathbf{K}^{-1}\mathbf{Y}'(\mathbf{x} - \gamma_1\,\mathbf{t}_1) \qquad (10)$$

where γ_1 is a constant to make $\mathbf{t}_2$ **R**-orthogonal to $\mathbf{t}_1$. After analogous extractions of *s* components the fitted values are

$$\hat{x}_s = \alpha_0 + \alpha_1\,\mathbf{t}_1 + ... + \alpha_s\mathbf{t}_s = \mathbf{R}^{-1}\,\mathbf{Y}\,\mathbf{u}^* \qquad (11)$$

in which the coefficients $\{\alpha_j\}$ are obtained by a regression in the **R**-norm of **x** on $\mathbf{t}_1, \mathbf{t}_2, ... , \mathbf{t}_s$ and

$$\mathbf{u}^* = \alpha_0 + \alpha_1\,\mathbf{u}_1 + ... + \alpha_s\,\mathbf{u}_s. \qquad (12)$$

Equations (11) and (12) show that the WA-PLS predictor is still a weighted average, but one with an updated weight vector. If the site totals are constant (**R** = c**I**), the distinction between PLS and WA-PLS lies in the estimator for the weight vector only.

It is interesting to note that the WA-PLS predictor lies in the space spanned by the vectors

$$\mathbf{A}\mathbf{x}, \mathbf{A}^2\mathbf{x}, ..., \mathbf{A}^s\mathbf{x} \qquad (13)$$

where **A** is the reciprocal averaging operator $\mathbf{R}^{-1}\ \mathbf{Y}\ \mathbf{K}^{-1}\ \mathbf{Y}'$. This result can be shown by expanding the components in terms of **x** (see Stone and Brook 1990 for a PLS equivalent) and shows the relation between WA-PLS and the reciprocal averaging algorithm of correspondence analysis (Hill 1982; Greenacre 1984).

Table 1 gives an example of the cross-validatory choice of the number of components using simulated data with 100 sites and 60 species. With each additional WA-PLS component, the root mean squared error (RMSE) steadily decreases, but at the danger of overfitting. Cross-validation by leave-one-out shows that the standard error of prediction (SEP) is minimal at 7 components (Table 1). Thus, the optimal number is 7. When the resulting transfer function is applied to an independent test set of 500 sites, it gives a slightly higher SEP than the leave-one-out estimate. The last column of Table 1 shows that usage of only 5 components would have given an even lower SEP, but in real applications of the method this column is not available.

4.7 **RIVAL METHODS**

Weighted averaging method (WA)

The weighted averaging method (Ter Braak and van Dam 1989; Birks *et al.* 1990a), a precursor of WA-PLS, is based on the idea that species' abundance or probability of occurrence is a unimodal function of the environmental variable to be calibrated. The method consists of three parts: WA regression, WA calibration and an additional 'deshrinking' regression. The parts are motivated as follows. A species with a particular optimum will be most abundant in sites with *x*-values close to its optimum. This motivates WA regression: estimate species optima simply by weighted averaging of the *x*-values of the sites, *i.e.* by (7) (Ter Braak 1985; Ter Braak and Looman 1986). Species present and abundant in a particular site will tend to have optima close to its *x*-value. This motivates WA calibration: estimate the *x*-values of the sites by weighted averaging of the species optima, *i.e.* by (8) (Ter Braak and Barendregt 1986). Because averages are taken twice, the range of the estimated *x*-values in $\mathbf{t}_1$ is shrunken. The amount of shrinking can be estimated from the training set by regression, *i.e.* by (9). Therefore, the weighted averaging method is equivalent to 1-component WA-PLS. There is a quibble attached to this statement. In WA, there are two variants of the deshrinking regression, either an 'inverse' regression ($\mathbf{x}$ on $\mathbf{t}_1$) or a regression ($\mathbf{t}_1$ on $\mathbf{x}$) (Birks *et al.* 1990a). Inverse regression minimizes the mean squared error in the training set, but at the cost of introducing bias at the endpoints. These issues are fully reviewed for linear calibration by Osborne (1991). WA-PLS uses the inverse regression. Thus, WA with inverse regression is equivalent to 1-component WA-PLS. WA-PLS shares the basic idea of WA, but with updated optima (12).

Gaussian model and multinomial logit model (MLM)

Calibration by maximum likelihood requires a response model. We first consider the systematic part of the model. A simple and tractable unimodal model for the expected abundance is the Gaussian model (*e.g.* Ihm and Van Groenewoud 1984), namely

$$\mathrm{f}_k(x) = \exp\,(c_k - ½\,(x - u_k)^2/w_k^2) \qquad (14)$$

where $\mathrm{f}_k(x)$ is the expected value of the abundance of species k ($k = 1...m$) as a function of the environmental variable x, and u_k, w_k and c_k are species parameters called the optimum, the tolerance (niche width) and the (log)maximum, respectively. As mentioned above, the vector of weighted averages $\mathbf{u}_1 = (u_{11}, ..., u_{1k}, ... u_{1m})'$ in (7) can be viewed as a simple estimator of the optima-vector of species, hence the usage of the symbol u in (14). Model (14) is not complete for compositional data because of the constraint that the abundances of all species per site sum to 1. If we take equation (14) as a model for absolute abundances of species at a site, then a model for expected proportions, E y_k, follows by division by their sum (Ihm and Van Groenewoud 1984)

$$\mathrm{E}\, y_k = \exp(\eta_k) \,/\, \{ \Sigma_{k'} \exp(\eta_{k'}) \} \tag{15}$$

with, in the terminology of GLM (McCullagh and Nelder 1989), a linear predictor that is a parabola in x,

$$\eta_k = c_k - ½\, (x - u_k)^2/w_k^2. \tag{16}$$

For equal tolerances ($w_k = w$), the linear predictor can be simplified considerably to

$$\eta_k = a_k + b_k\, x, \tag{17}$$

with $a_k = c_k$ - ½ u_k^2/w^2 and $b_k = u_k/w^2$; the quadratic term in x cancels out because it occurs in both the numerator and denominator of (15). We restrict attention to the equitolerance case in the parametrization of (17), because it combines unimodality (Anderson 1984), parsimony of parameters and computational tractability. The relation with Goodman's (1986, 1991) family of RC-association models follows by noting that

$$\log(\mathrm{E}\, y_{ik}) \;= \varphi_i + a_k + b_k x_i \tag{18}$$

with $\varphi_i = -\log(\Sigma_k \exp(\eta_{ik}))$. It is a constrained association model, because x_i is known in the training set.

For the stochastic part of the response model a multinomial distribution is assumed. Then, model (15) with (17) is known under the name of multinomial logit model (MLM) and can be fitted by maximum likelihood through fitting (18) by log-linear regression (McCullagh and Nelder 1989: pp. 211-212). This results in estimates of the species parameters $\{a_k, b_k\}$. An unknown x_0 is estimated from the species data of a site by loglinear regression with the $\{b_k\}$ as values of the predictor variable and the $\{a_k\}$ as offset (*i.e.* a predictor with a unit regression coefficient). Sampling errors in $\{a_k, b_k\}$ are ignored. WA can be considered as an approximation to this maximum likelihood method (Ter Braak 1988; Ter Braak and van Dam 1989). Classical calibration sometimes benefits from some form of shrinkage. We study the usefulness of an extra regression of $\{x_i\}$ on $\{_i\}$.

5 **METHODS**

6.1 **SIMULATIONS**

The training set consists of 100 sites arranged in an L-shape with respect to two environmental variables, a calibration variable (x) and a nuisance variable (z) (Fig. 1): 10 × 5 sites on a rectangular lattice in the area [0, 100] × [0, 25] and 5 × 10 sites in the area [0, 50] × [0, L_z]. The effect of z studied by varying L_z from 100 (L-shape, Fig. 1) to 50 and 25. The training set was designed by analogy with typical water chemistry data of pH and Al, in which Al is more variable at low pH than at high pH and in which there are more acid than alkaline sites.

The species compositional data were generated with COMPAS (Minchin 1987) in four steps. First, unimodal response surfaces with respect to x and z were generated that randomly vary in skewness and kurtosis among species. Secondly, the expected abundance of the k-th species at the i-th site, μ_{ik}, was read off the unimodal response surface of the k-th species at the coordinates (x_i, z_i). Thirdly, a count was sampled from a Poisson distribution with expected value μ_{ik}. Finally, the counts so derived were expressed as percentages of the total count per site.

The details of the generation of the response surfaces are as follows. In COMPAS, a species response surface is the product of two unidimensional curves where each curve is a unimodal, generalized beta function with four parameters, which vary randomly between species, independently for x and z. They are: the range (r) over which the curve rises and falls, the modal coordinate or optimum (u) and two shape parameters, α and γ, that govern the skewness and kurtosis. A final parameter is the maximum (A_u) at the modal coordinates. The length of gradient can then be defined in Range (R) units as (environmental range)/$\bar{r}$, where $\bar{r}$ is the mean range of the species curves. The gradient lengths of x and z, denoted by R_x and R_z, are varied between simulations. If $R \leq 0.5$, most curves appear as being monotonic increasing or decreasing along the sampled interval, whereas if $R \geq 1$, most curves have a clear optimum in the sampled interval. The following settings were used: r is uniform between 50/R and 150/R, u is uniform between -75/R and 175/R (thus spaciously embracing the sampling interval), α and γ are independent uniform between 0.5 and 3.5 (giving skew and flat-topped curves) and A_u is lograndom between 10 and 80 (*i.e.* log(A_u) is uniform on the interval [log(10), log(80)]).

Three series of comparisons were carried out. In the first series WA-PLS is compared with WA with special attention paid to the effect of the nuisance variable. If z influences the species composition more, the signal-to-noise ratio for calibrating x decreases, giving an increase in the prediction error. WA-PLS is designed to exploit the structure in the noise. It is thus expected that the relative performance of WA-PLS to WA will increase with the influence of z. In the second series WA-PLS is compared with autoscaled PLS with and without logarithmic transformation of the species data (ln(y_{ik} +1)). Attention focuses here on the effect of the lenght of gradient of x on the relative performance of WA-PLS. We expected that the relative performance of WA-PLS to PLS will increase with length of gradient. For the smallest value of R studied (R = 0.5), PLS is expected to outperform WA-PLS. In the third series, WA-PLS is compared with MLM. Because the nuisance variable is unknown in practice, we fit the model without it. Because of the multinomial assumption, the method disregards residual structure in the species data after fitting x. We thus expect WA-PLS to outperform MLM if z is influencial.

In the first simulation series, WA-PLS and WA were applied to 54 training sets. Underlying the 54 sets is a split-plot 2×3^2 experiment (Cochran and Cox 1957) with the number of species (m) and the length of the calibration gradient R_x on whole plots and the influence of z (expressed as gradient length R_z) on subplots. A whole plot is one independently simulated set of species parameters across the (x,z)-plane. Each whole plot is divided in three subplots by varying L_z. Species data sets were generated independently for each subplot. The levels of the factors are: $m = 60, 150$; $R_x = 0.5, 1.0, 2.0$; $R_z = 0.25, 0.5, 1.0$. R_x is adjusted by varying the average species range, while keeping the range of x equal to 100. R_z is adjusted by varying L_z ($L_z = 25$, 50 and 100), while keeping the average species range equal to 100. All factor combinations with $m = 60$ were replicated four times, those with $m = 150$ twice.

The second simlation series consists of the data of series 1, but is restricted to $L_z = 100$. The two variants of PLS were carried out on these 18 data sets. In the third series MLM was applied to the 8 data sets of series 1 with $m = 60$, $R_x = 1$ and $R_z = 0.25$ and 1.

The performance of the calibration methods is measured by the standard error of prediction (SEP), defined as the root mean squared prediction error and estimated either from the training set by leave-one-out (*e.g.* Hastie and Tibshirani 1990) or from an independent test set. Three test sets were used. Test set A consists of 500 sites arranged on a 100 × 5 rectangular lattice in the area [0, 100] × [0, 25]. This is the area that is common to all training sets. Test set B also consists of 500 sites and is a five-times replication of the L-shaped training set. Test set B is used in the second series of comparisons in which all training sets have this shape. Test set C consists of 125 sites arranged on a 25 × 5 rectangular lattice in the area [50, 75] × [30, 100]. There are no training sites in this area. With test set C we explore the danger of hidden extrapolation relevant to the no-modern analogue problem in palaeoenvironmental reconstructions.

6.3 **REAL DATA**

The methods were also applied to the Imbrie and Kipp (1971) data set collected to infer summer and winter sea-surface temperature (SST) and salinity from foraminifera in the CLIMAP program. The training set consists of 61 core top samples, most of them taken in the Atlantic Ocean between latitudes 60°N and 54°S, and contains 27 species of foraminifera. The abundance of each species is expressed as a percentage of the total count per core-top sample. Further details and the full data can be found in Imbrie and Kipp (1971).

6.5 **DATA ANALYSIS**

In the training sets the number of species present in a site was always below 50. For weighted averaging methods, the number of species in a site is, however, not an appropriate statistic. For example, if there are three species present with abundances 100, 1, and 1, respectively, the first species takes nearly all the weight so that the effective number of species is close to 1. A good measure for this effective number is Hill's (1973a) N_2 measure of diversity, which is the reciprocal of Simpson's diversity index. We report the median and range of the effective numbers of species per site and of the effective number of occurrences per species (defined analogously). To further characterize the training sets in data-analytical terms, we carried out a detrended canonical correspondence analysis using the program CANOCO 3.1 (Ter Braak 1987-1990) with *x* as the only environmental variable. Detrending-by-segments was used. The length of gradient of the first axis of this analysis is the gradient length of *x* in SD-units (Hill and Gauch 1980). We also report the gradient length of the second, unconstrained axis, the first two eigenvalues (λ_1 and λ_2) and the percentage variance explained by the constrained axis (100 * λ_1/total inertia).

PLS and WA-PLS were carried out using the program CALIBRATE (Juggins and ter Braak 1992). The Gaussian model in parametrization (18), *i.e.* MLM, was implemented using GENSTAT 5 (1987). The simulation results were analyzed by analysis of variance (ANOVA) with GENSTAT 5. Standard errors of prediction were transformed to natural logarithms prior to the analysis. Results are reported after backtransformation, i.e. as geometric means and coefficients of variation (CV) (Aitchison and Brown 1969). The standard error of differences of means of the ANOVA is backtransformed similarly to the coefficient of variation of a ratio (CVR). The tables of means that we report, contain significant effects as judged on the basis of the usual F-test at the 5% level. The relative performance of WA-PLS with respect to a method M, say, is expressed as the geometric mean of the ratio of the SEP of M to that of WA-PLS.

7 **RESULTS**

8.1 **SIMULATIONS**

Data summary

Fig. 2 shows typical examples of the relationship of species with the calibration variable *x* in the simulations with $R_x = R_z = 1$. With $R_x = 1$, the unimodal curves span, on average, 100 *x*-units, as exemplified by Fig. 2b which rises and falls approximately between *x*= 10 and 110, but can be partly (Fig. 2d) or fully (Fig. 2a) truncated depending on the position of the mode of the response surface. The truncation at *x* = 50 in Fig. 2c is an effect of the L-shaped site configuration; the mode of the species in Fig 2c lies at coordinates (78, 63) in Fig. 1. Thus, the scatter is due to both the nuisance variable *z* and the random noise.

Tables 2 and 3 summarize the training sets in data-analytical terms. If the gradient lengths (R_x and R_z) increase, the average curve width decreases. Each curve then covers fewer sites so that the effective number of occurrences per species (mean *ca.* 20) decreases (Table 2a,b). In the simulations with 60 species (*m* = 60), the effective number of species per site is on average *ca.* 6. It depends most on R_x (Table 2c); the effect of R_z is minor. For m = 150, the effective number is about twice as large. As expected, the constrained first eigenvalue of the canonical correspondence analysis increases with R_x, the unconstrained second eigenvalue increases with R_z (Table 3). The length of the first gradient in SD-units increases with R_x from *ca.* 2 SD to 8 SD. For m = 150, the length is on average 0.4 SD higher than for m = 60. The percentage variance explained by the calibration variable varies among training sets between 13 and 41%.

Simulation series 1: comparison of WA-PLS and WA

Fig. 3 shows for the example simulation data set of Table 1 and Fig. 2 how WA-PLS reduces the residuals. After extraction of the first component (*i.e.* in WA), the sites with large values of the nuisance variable (z) and average values for the calibration variable (x) have large negative residuals. Extraction of further components reduces the residuals and the structure therein (Figs 3b-3d). From Table 1, the optimal number of components is estimated as 7, giving a relative performance of WA-PLS with respect to WA of 2.37 (= 5.94/2.51) in the training set and of 2.14 (= 5.68/2.65) in test set B.

In simulation series 1, the SEP of WA-PLS as estimated by leave-one-out varies between 1.0 to 5.5 among the training sets, the geometric mean being 2.5. The SEP is thus, at most, 5% of the range of the calibration variable. The effects of the design variables, as detected by ANOVA, are as follows. The SEP decreases by a factor of *ca.* 2 with both the number of species and the length of the calibration gradient (Tables 4a and 4c). The length of the nuisance variable has a minor effect (< 12%) (Table 4b). The optimal number of components varies between 2 and 10 (the maximum, attained once, in our simulations) with an average of 6. WA (*i.e.* 1-component WA-PLS) thus never performed best as assessed by the SEP. For a short nuisance gradient the optimal number is two smaller than for a long one ($P<0.03$). The optimal number is quite variable. The coefficient of variation among replications with the same species parameters is 30%.

The relative performance of WA-PLS with respect to WA as based on their SEP's is about 2.0 and was influenced little by the length of the calibration gradient. The length of the nuisance gradient had little effect on the relative performance as well, except for a short calibration gradient. If R_x is 0.5, the relative performance increases with R_z (Table 4d). Similar values and trends were found in the test set A. The poor performance of WA for a short calibration and a long nuisance gradient (R_x = 0.5 and R_z = 1.0) is largely due to bias. In this instance the bias in the WA-predictor in the interval [0,25] in test set A is a large as 11!

With test set C the prediction error in no-analogue situations (extrapolation) is explored. The extrapolation is less severe for R_z = 1.0. The standard error of prediction decreases correspondingly with R_z from 14 to 7 (Table 5). The predictions for set C may thus be unacceptably bad, thus highlighting the importance of the no-analogue problem. In the very bad cases (R_z = 0.25), WA-PLS performs worse than WA (rp = 0.63). If R_z = 0.5 and 1.0, the relative performances are on average 1.16 and 2.23, respectively, and decrease with R_x. It appears that WA-PLS deteriorates quicker than WA with increasing extrapolation.

Simulation series 2: comparison of WA-PLS and PLS

The comparison is based on all L-shaped training sets with R_z = 1.0 and test sets B and C.

The variants of PLS (with and without logarithmic transformation) did not differ significantly in performance. We report the relative performance with respect to untransformed PLS only. In test set B, WA-PLS outperformed PLS in all cases, except one in which the SEP's were about equal. The geometric mean of the relative performances of WA-PLS compared to PLS is 1.33. The relative performance increases with the length of the calibration gradient and the number of species (Table 6). The optimal number of components did not differ much between methods. WA-PLS performed marginally better under extrapolation: the relative performance in test set C was 1.15 (CV = 8%), which is just significantly different from 1 at the 10% level. The length of the calibration gradient and the number of species have no significant effect on the relative performance in test set C.

It is interesting to see how far the SEP, as estimated from the training set by leave-one-out, is a good estimator of the real SEP. For this, we compared the leave-one-out SEP with the SEP as estimated from test set B (which has the same distribution as the training set). To our initial surprise, the leave-one-out estimator gave an overestimation of 7% of the real SEP (CV = 3%, $P<0.05$) for both WA-PLS and PLS. We could have known better: Hastie and Tibshirani (1990) proved that the leave-one-out SEP is overpessimistic for all biased regression methods. The overpessimism is apparently greater than the optimism introduced by the cross-validatory choice of the optimal number of components.

Simulation series 3: comparison of WA-PLS and MLM

Assessed by the SEP in test set A, WA-PLS outperforms MLM for a long nuisance gradient ($R_z = 1$), and performs about equally well if the nuisance gradient is short (Table 7). The extra inverse regression step in MLM improves the results a little. With extrapolation (set C), WA-PLS performs significantly better if $R_z = 1$, but worse ($P<0.07$) if $R_z = 0.25$. We experienced convergence problems in prediction with MLM. A good initial estimate is needed for which we used perturbated true values (in practice one could use the WA-PLS estimate). In one case of $R_z = 0.25$, the iterations nevertheless diverged for some sites in test set C. This case was treated as a missing value in Table 7b.

8.3 **REAL DATA**

Table 8 summarizes the Imbrie and Kipp (1971) data. The effective numbers of species per site are similar to those in the simulation data. The gradient length of *ca.* 4 SD of summer and winter sea-surface temperature (SST) demonstrates the unimodal nature of the abundance data. For salinity the gradient length is smaller. In comparison with the simulation data, the secondary gradient is short. Summer and winter SST are highly correlated (r = 0.97) and their correlation with salinity is *ca.* 0.7. A detrended canonical correspondence analysis of the species with all three as explanatory variables gives the eigenvalues 0.75, 0.13 and 0.01. The fourth, unconstrained axis has a small eigenvalue ($\lambda_4 = 0.05$). These results demonstrate the essential one-dimensionality of these data. Thus, seasonality (Summer - Winter SST) does not have a strong signal in the abundance data.

Table 9 shows the RMSE and, if available, the leave-one-out SEP for various methods. The optimum number of components in WA-PLS is 3 for all three variables. On the basis of the

SEP, the relative performance of WA-PLS with respect to WA is 1.22, 1.48 and 1.13 for summer SST, winter SST and salinity, respectively. Fig. 4 shows the predicted values and prediction residuals for winter SST after 1 (WA) and 3 components. The curvature in Fig 5a (WA) is largely removed after extraction of three components. The bias that WA gives in the range 3 - 15°C is diminished after extracting 3 components. The residuals after three components still show structure, but taking more than three components increases the prediction error. The plot of the first two components of WA-PLS (Fig. 5) shows a parabolic relation that is well known from correspondence analysis (arch or Guttman effect; Greenacre 1984). Apparently the arch effect is exploited in WA-PLS to improve the prediction. The optimum number of components in PLS is higher than in WA-PLS for summer SST and winter SST. Despite the low RMSE, the SEP of PLS is higher than that of WA-PLS, most notably for winter SST. Despite its sophistication, MLM performs slightly worse than WA-PLS.

For the remaining six methods, applied by other authors, we have no leave-one-out SEP's available. This complicates the comparison with the previous methods. Following Ter Braak and van Dam (1989), Birks (unpubl.) fitted, for all species, Gaussian logit curves to each calibration variable in turn. With respect to winter SST, for example, 21 of the 27 species showed a significant optimum, 3 species showed a significantly increasing sigmoid logit curve, one species a significantly decreasing curve and 2 species showed no significant relation. Subsequently, winter SST was reconstructed for each training site from its abundance data by maximum likelihood with use of the fitted curves (with the help of a special-purpose numerical optimization program incorporated in WACALIB; Line and Birks 1990) and by weighted averaging with use of the estimated optima or estimated optima and tolerances (Ter Braak and van Dam 1989, Birks *et al.* 1990a). The latter method, WA_{tol} calibration, uses the squared tolerance to downweight a species with a large curve width (more specifically, in the weighted average (17), each abundance value y_{ik} is replaced by y_{ik}/w_k^2). Surprisingly, the ML calibration variant is outperformed by the weighted averaging variants. None of them is expected to outperform WA-PLS in terms of SEP. Imbrie and Kipp (1971) applied a form of principal components analysis, called Q-mode factor analysis (QFA; Klovan and Imbrie 1971), to the abundance data and regressed the calibration variables on to the first four components in both a linear and a quadratic model. Roux (1979) replaced the factor analysis by correspondence analysis and used forward selection in the subsequent linear regression on the components. The RMSE's for these methods are corrected for the number of parameters in the regression phase of the model, but disregard the random nature of the components and the selection process. For these data, Roux's correspondence analysis regression outperforms Imbrie and Kipp's method and ranks with WA-PLS among the best. Being an ancestor to WA-PLS this is no surprise, especially for a data set in which the first correspondence analysis axis is so highly correlated with the calibration variables.

9 **DISCUSSION**

The simulations and the real data show that, for compositional data, WA-PLS is a better species-environment calibration method than either WA or PLS. It performs equally well or, if there are secondary gradients, better than MLM, the maximum likelihood method based on the equitolerance Gaussian model for multinomial data. In cases where no modern analogues exist (hidden extrapolation), all methods perform poorly and no uniform winner emerges. For very strong extrapolation, WA may perform better than WA-PLS.

We expected that PLS would outperform WA-PLS for short gradients. But the simulations show good a performance for WA-PLS even for the shortest gradient length tested (0.5 Range units, *ca.* 2 SD units). Apparently, WA-PLS is well-suited for compositional data.

Another expectation was that the relative performance of WA-PLS to simple WA would increase with the importance of the nuisance gradient. In the simulations, however, this relation was barely visible, except for short primary gradients. In hindsight, the result is not so surprising. The expectation was formulated on the assumption that WA-PLS would not improve WA in a system governed by a single environmental variable. But, this assumption proved to be false. In one-dimensional simulations (Ter Braak and Juggins 1993), WA-PLS also improves WA in that it corrects the edge effects (and related non-linear distortions) that occur in WA. In these cases, both WA-PLS and correspondence analysis yield components that are polynomials of the first (*e.g.* Fig. 5), an phenomenon known as the Guttman effect (Greenacre 1984) or arch effect (Hill and Gauch 1980). The higher components achieve a non-linear rescaling of the first component, *i.e.* of the simple weighted average estimate. Apparently, Ter Braak and van Dam (1989) were wrong in believing that those 'spurious' axes of a correspondence analysis were useless for prediction. However, in noisy data sets little can be gained. Ter Braak and Juggins (1993) varied the amount of qualitative noise in simulations by randomly replacing abundance values by zeroes. If the probability thereof is increased to 75%, the gain of WA-PLS over WA vanishes. Only if the noise is highly structured, as in our simulations, does the gain of WA-PLS over WA continue.

It might be useful to combine (non)linear regression and WA-PLS. In pollution studies, a major part of the information could come from the total abundance and number of species. In other studies one might ask whether one needs the full species assemblage for calibration or whether a summary statistic, like the diversity or total biomass, would suffice. Such problems ask for a combination of regression and WA-PLS. The natural approach to this problem is first to fit the regression followed by a WA-PLS in which the components are restricted to be orthogonal to the regressor variables of the first step. In this approach it would be unnatural if the implicit site weights would differ among the steps. With unequal weights, what would orthogonality mean?. One therefore needs to make the weights equal, preferably, by transforming the WA-PLS data to compositions.

Would knowledge of other environmental variables than the one of interest have helped to reduce the prediction error? In practice such nuisance variables are typically known in the training set only and should be estimated during reconstruction. This requires simultaneous calibration of several variables by classical multivariate calibration. Brown (1982) and Lorber *et al.* (1987) argue that this has no advantage over inverse regression for multivariate linear models with normal error. But does this carry through to nonlinear, unimodal models with Poisson or multinomial error? In attempts to develop practical methods for joint calibration based on the

Gaussian model (14), we considered the utility of canonical correspondence analysis, which is an approximation to this model (Ter Braak 1986, 1988). Classical calibration with (full-rank) canonical correspondence analysis amounts to a multivariate linear regression of the columns of $\mathbf{Y}^*$ (5) on the environmental variables transformed as $\mathbf{x}^*$ in (5). Turning this classical calibration model upside down leads to an inverse regression of $\mathbf{x}^*$ on $\mathbf{Y}^*$, but would fail because of multicollinearity problems. WA-PLS is the inverse regression approach that guards against this. It remains to be seen whether joint calibration has advantages in the original Gaussian model (14) - (18).

We see no advantage in using the multivariate version of PLS or WA-PLS (PLS2 and WA-PLS2, respectively) for joint calibration of environmental variables; it is a form of inverse regression, albeit multivariate, instead of being a form of multivariate classical calibration. Perhaps WA-PLS2 has advantages in multiple discriminant problems that use species composition data.

As the simulations show, the theoretical advantages of maximum likelihood calibration based on (18) are outweighed by the disadvantages of the assumptions of a multinomial distribution and the rigid functional form of the response functions. The development of generalized linear mixed models and generalized estimation equations (Shall 1991; Engel and Keen 1992) may alleviate the multinomial assumption and allow the use of the residual correlations among species caused by unknown nuisance variables. Moreover, spline models (Hastie and Tibshirani 1990) may allow less rigid functional forms. Until the time that such sophisticated methods mature and demonstrate their power for species-environment calibration, WA-PLS is recommended as a simple and robust alternative.

11 **ACKNOWLEDGEMENTS**

We would like to thank Margriet Stapel and Ajos A. M. Jansen for comments. Part of the work reported here was conducted while S.J. was in receipt of a Royal Society Fellowship (UK).

13 **REFERENCES**

Aitchison, J. & J. A. C. Brown, 1969. The Lognormal Distribution. Cambridge University Press, Cambridge.

Anderson J. A., 1984. Regression and ordered categorical variables. J. R. Statist. Soc. B 46: 1-30.

Bartlein P. J. & T. Webb, 1985. Mean July temperature at 6000 yr B P. in Eastern North Ammerica: regression equations from fossil-pollen data. Syllogeus 55: 301-342.

Battarbee R. W., 1984. Diatom analysis and the acidification of lakes. Phil. Trans. R. Soc. Lond. B 305: 451-477.

Birks H. J. B., J. M. Line, S. Juggins, A. C. Stevenson & C. J. F. Ter Braak, 1990a. Diatoms and pH reconstruction. Phil. Trans. R. Soc. Lond. B 327: 263-278.

Birks, H. J. B., S. Juggins & J. M. Line, 1990b. Lake surface-water chemistry reconstructions from palaeolimnological data. In: The Surface Waters Acidification Programme (Ed. B. J. Mason), pp. 301-313. Cambridge university Press, Cambridge.

Brown P. J., 1982. Multivariate calibration. J. R. Statist. Soc. B 44: 287-321.

Cumming B. F., J. P. Smol & H. J. B. Birks, 1991. The relationship between sedimentary chrysophyte scales (Chrysophyceae and Synurophyceae) and limnological characteristics in 25 Norwegian lakes. Nord. J. Bot. 11: 231-241.

Cochran, W.G., Cox, G.M. 1957. Experimental designs (2nd Edition). Wiley (New York).

COHMAP Members, 1988. Climatic changes of the last 18,000 years: observations and model simulations. Science 241: 1043-1052.

Dixit S. S., A. S. Dixit & J. P. Smol, 1989. Relationship between chrosophyte assemblages and environmental variables in seventy-two Sudbury lakes as examined by canonical correspondence analysis (CCA). Can. J. Fish. Aquat. Sci. 46: 1667-1676.

Dixit S. S., A. S. Dixit & J. P. Smol, 1991. Multivariable environmental inferences based on diatom assemblages from Sudbury (Canada) lakes. Freshwater Biology 26: 251-266.

Dixit S. S., J. P. Smol, J. C. Kingston & D. F. Charles, 1992. Diatoms: powerful indicators of environmental change. Environmental Science and Technology 26: 23-33.

Engel B. & A. Keen, 1993. A simple approach for the analysis of generalized linear mixed models. Statist. Neerl. 47: 00-00.

Fritz S. C., S. Juggins, R. W. Battarbee & D. R. Engstrom, 1991. Reconstruction of past changes in salinity and climate using a diatom-based transfer function. Nature 352: 706-708.

Gasse F. & F. Tekaia, 1983. Transfer functions for estimating paleoecological conditions (pH) from East African diatoms. Hydrobiologia 103: 85-90.

GENSTAT 5 Committee, 1987. GENSTAT 5 Reference manual. Clarendon Press, London.

Goodman L. A., 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. Int. Stat. Rev. 54: 243-309.

Goodman L. A., 1991. Measures, models and graphical displays in the analysis of cross-classified data. J. Amer. Stat. Assoc. 86: 1085-1138.

Greenacre M. J., 1984. Theory and applications of correspondence analysis. Academic Press, London.

Hall R. I. & J. P. Smol, 1992. A weighted-averaging regression and calibration model for inferring total phosphjorus concentration from diatoms in British Columbia (Canada) lakes. Freshwater Biology 27: 417-434.

Hastie T. & R. Tibshirani, 1990. Generalized Additive Models. Chapman and Hall, London.

Hill M. O., 1973a. Diversity and evenness: a unifying notation and its consequences. Ecology 54: 427-432.

Hill M. O., 1973b. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61: 237-249.

Hill M. O., 1974. Correspondence analysis: a neglected multivariate method. Applied Statistics 23: 340-354.

Hill M. O., 1982. Correspondence analysis. Encyclopedia of Statistical Sciences 2: 204-210.

Hill M. O. & H. G. Gauch, 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio 42: 47-58.

Ihm P. & H. van Groenewoud, 1984. Correspondence analysis and Gaussian ordination. Compstat Lectures 3: 5-60.

Imbrie, J. & N. G. Kipp, 1971. A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Caribbean core. In: The late Cenozoic glacial ages (Ed. K. K. Turekian), pp. 77-181. Yale University Press, New Haven.

Jongman R. H. G., C. J. F. ter Braak & O. F. R. van Tongeren, 1987. Data analysis in community and landscape ecology. Pudoc, Wageningen.

Juggins S. & C. J. F. ter Braak, 1992. CALIBRATE - a program for species-environment calibration by [weighted-averaging] partial least squares regression. Environmental Change Research Centre, University College, London.

Klovan J. E. & J. Imbrie, 1971. An algorithm and FORTRAN-IV program for large scale Q-mode factor analysis and calculation of factor scores. Math. Geol. 3: 61-67.

Line J. M. & H. J. B. Birks, 1990. WACALIB version 2.1 - a computer program to reconstruct environmental variables from fossil assemblages by weighted averaging. J. Paleolimn. 3: 170-173.

Lorber A., L. E. Wangen & B. R. Kowalski, 1987. A theoretical foundation for the PLS algorithm. J. Chemometrics 1: 19-31.

Martens H. & T. Naes, 1989. Multivariate calibration. Wiley, Chichester.

McCullagh P. & J. A. Nelder, 1989. Generalized linear models (2nd Edition). Chapman and Hall, London.

Minchin P. R., 1987. Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio 71: 145-156.

Odum E. P., 1971. Fundamentals of Ecology (3rd Edition). W.B. Saunders Company, Philaelphia.

Oksanen J., E. Laara, P. Huttunen & J. Merilainen, 1988. Estimation of pH optima and tolerances of diatoms in lake sediments by the methods of weighted averaging, least squares and maximum likelihood, and their use for the prediction of lake acidity. J. Paleolimn. 1: 39-49.

Osborne C., 1991. Statistical calibration: a review. Int. Statist. Rev. 59: 309-336.

Roux M., 1979. Estimation des paléoclimats d'apres l'ecologie des foraminifès. Cahiers de l'Analye des Données 4: 61-79.

Shall R., 1991. Estimation in generalized linear models with random effects. Biometrika 78: 719-728.

Sloan L. C. & E. J. Barron, 1992. A comparison of Eocene climate model results to quantified paleoclimate interpretations. Palaeogeography, Palaeoclimatology, Palaeoecology 93: 183-202.

Spellerberg I. F., 1991. Monitoring ecological change. Cambridge University Press, Cambridge.

Stone M. & R. J. Brooks, 1990. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. J. R. Statist. Soc. B 52: 237-269.

Ter Braak C. J. F., 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal reponse model. Biometrics 41: 859-873.

Ter Braak C. J. F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67: 1167-1179.

Ter Braak C. J. F., 1987-1990. CANOCO - a FORTRAN program for CANOnical Community Ordination. Microcomputer Power, Ithaca, NY, USA.

Ter Braak, C. J. F., 1988. Partial canonical correspondence analysis. In: Classification and related methods of data analysis (Ed. H. H. Bock), pp. 551-558. North-Holland, Amsterdam.

Ter Braak C. J. F. & L. G. Barendregt, 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. Math. Bio. 78: 57-72.

Ter Braak C. J. F. & S. Juggins, 1993. Weighted averaging partial least squares regression (WA-

PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia (in press).

Ter Braak C. J. F. & C. W. N. Looman, 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio 65: 3-11.

Ter Braak C. J. F. & H. van Dam, 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia 178: 209-223.

Walker I. R., R. J. Mott & J. P. Smol, 1991. Allerød - Younger Dryas lake temperatures from midge fossils in Atlantic Canada. Science 253: 1010-1012.

Walker I. R., J. P. Smol, D. R. Engstrom & H. J. B. Birks, 1991. An assessment of Chironomidae as quantitative indicators of past climatic change. Canadian Journal of Fisheries and Aquatic Sciences 48: 975-987.

Whittaker R. H., S. A. Levin & R. B. Root, 1973. Niche, habitat and ecotope. Amer. Natur. 107: 321-338.

Whitton B. A. & E. F. S. Rott, 1991. Use of algae for monitoring rivers. Inst. Botanik, Wien.

Wold S., A. Ruhe, H. Wold & W. J. Dunn III, 1984. The collinearity problem in linear regression: the partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5: 735-743.

15 **APPENDIX**

THEOREM: The solution of WA-PLS can be obtained from a PLS program by preprocessing **Y** and **x** by (5) for an **R**-centred **x** and by postprocessing the results by (6).

For the proof we need the

LEMMA: The centring constraint on **u** in criterion (4) of WA-PLS is automatically fulfilled by **R**-centring **x** as a preprocessing step.

PROOF of the LEMMA: For an **R**-centred **x** ($\mathbf{1}'\mathbf{R}\,\mathbf{x} = 0$), $\mathbf{t}'\mathbf{R}\mathbf{x}$, the numerator of the criterion (4), is invariant under translation of $\mathbf{t} = \mathbf{R}^{-1}\mathbf{Y}\mathbf{u}$, and thus under translation of the weights **u**, whereas the minimum of the denominator is attained for centred **u**; hence the criterion (4) is maximized for centred **u**.

PROOF of the THEOREM: The proof is by inserting the transformation (5) in criterion (2) of PLS and rewriting it in the form of criterion (4) of WA-PLS. For the denominator of (2) we have, with (6), $\mathbf{b}'\mathbf{b} = \mathbf{u}'\,\mathbf{K}\,\mathbf{u}$ and for the numerator, with (5) and (6),

$$(\mathbf{Y}^{*}\,\mathbf{b})'\,\mathbf{x}^{*} = (\mathbf{R}^{-½}\,\mathbf{Y}\,\mathbf{K}^{-½}\,\mathbf{b})'\,\mathbf{x}^{*} = (\mathbf{R}^{-1}\,\mathbf{Y}\,\mathbf{u})'\,\mathbf{R}\,\mathbf{x}$$

which shows the equivalence, except for the extra constraint, $\mathbf{1}'\mathbf{K}\,\mathbf{u} = 0$. But, from the lemma, the constraint in (4) is automatically taken care of with an **R**-centred **x**. Finally, because of the pre- and postprocessing steps (5) and (6) an inner product in PLS translates into an **R**-weighted inner product; for example, for components i and j, with $\mathbf{u}_i = \mathbf{K}^{-½}\,\mathbf{b}_i$ and $u_j = \mathbf{K}^{-½}\,\mathbf{b}_j$

$$(\mathbf{Y}^{*}\,\mathbf{b}_i)'\,(\mathbf{Y}^{*}\,\mathbf{b}_j) = (\mathbf{R}^{-1}\,\mathbf{Y}\,\mathbf{u}_i)'\,\mathbf{R}\,(\mathbf{R}^{-1}\,\mathbf{Y}\,\mathbf{u}_j).$$

Therefore, orthogonality of the linear combinations in PLS translates into **R**-orthogonality of the corresponding weighted averages in WA-PLS, as required. This completes the proof.

Table 1. Effect of the number of components (*s*) on the apparent error (RMSE) and prediction error (SEP) of WA-PLS in a simulated training set (n = 100, m = 60, $R_x = R_z = 1$) and test set (n = 500). The estimated optimum number of components is 7 because this number gives the lowest SEP, estimated by leave-one-out, in the training set. The last column is not available in real applications. For further explanation, see text.

	Training set		Test set B
s	RMSE	SEP	SEP
1	5.62	5.94	5.68
2	3.60	3.91	3.84
3	2.91	3.34	3.25
4	2.41	2.89	2.80
5	2.09	2.67	2.53
6	1.82	2.55	2.59
7	1.76	2.51 (*)	2.65
8	1.72	2.53	2.66
9	1.67	2.53	2.74
10	1.65	2.57	2.76

Table 2. Simulated training sets: the effect of the gradient lengths R_x (a) and R_z (b) on the effective number of occurrences per species (N_2) and (c) the effect of R_x on the effective number of species per site (M_2). Tabulated are means of the minimum (min), median (med) and maximum (max) of N_2 and M_2 per training set, in (c) restricted to the replications with 60 species (m = 60).

(a)

R_x	0.5	1.0	2.0
N_2min	1	1	1
med	20	13	12
max	86	71	44

(b)

R_z	0.25	0.50	1.0
N_2min	1	1	1
med	19	15	11
max	74	69	59

(c)

R_x	0.5	1.0	2.0
M_2min	5	3	1
med	8	6	5
max	12	10	9

Table 3. Simulated training sets: effect of R_x (a) and R_z (b) on the eigenvalue (λ) and the length of gradient in Standard Deviation units (SD) of the first and second axes (subscripts 1 and 2) of a detrended canonical correspondence analysis of the simulated species data with respect to *x*. V^2 is the percentage variance of the species data explained by the first two components. Tabulated are means with standard error of differences of means (sed).

(a)

R_x	0.5	1.0	2.0	sed
λ_1	0.35	0.68	0.89	(0.03)
SD_1	2.38	4.48	8.08	(0.17)
V^2	25.97	26.86	19.14	(1.51)

(b)

R_z	0.25	0.50	1.0	sed
λ_2	0.26	0.37	0.71	(0.03)
SD_2	2.22	2.86	5.21	(0.14)
V^2	31.37	24.43	16.18	(0.54)

Table 4. Simulation series 1: the effect of R_x (a), R_z (b) and m (c) on the SEP of WA-PLS in the training sets and (d) the effect of R_z on the relative performance (rp) of WA-PLS with respect to WA in the training sets with R_x = 0.5. If R_x = 1 or 2, rp ≈ 2.0, irrespective of R_z. (CV(R): 100 times the coefficient of variation (of a ratio) of the table entries).

(a)

R_x	0.5	1.0	2.0	CVR	CV
SEP	3.7	2.4	1.8	8	6

(b)

R_z	0.25	0.50	1.0	CVR	CV
SEP	2.4	2.6	2.7	4	4

(c)

m	60	150		CVR	CV
SEP	3.0	1.7		8	5

(d)

R_z	0.25	0.50	1.0	CVR	CV
rp w.r.t. WA	1.74	2.06	2.93	9	10

Table 5. Simulation series 1: effect of R_z on the SEP of WA-PLS in test set C (extrapolation) and the corresponding relative performance with respect to WA.

R_z	0.25	0.50	1.0	CVR	CV
SEP	14.1	11.5	6.6	18	15
rp w.r.t. WA	0.63	1.16	2.23	20	14

Table 6. Simulation series 2 (R_z = 1): effect of R_x (a) and m (b) on the SEP of WA-PLS in test set B and the corresponding relative performance (rp) with respect to PLS. CV and CVR are about 7 and 10%, respectively.

	(a)			(b)	
	R_x			m	
	0.5	1.0	2.0	60	150
SEP	3.60	2.40	1.78	3.10	1.60
rp w.r.t. PLS	1.21	1.23	1.60	1.24	1.54

Table 7. Simulation series 3 (m = 60, R_x = 1): effect of R_z on the SEP of WA-PLS in test sets A (a) and C (b) and the corresponding relative performance (rp) with respect to the MLM with (+R) and without an extra inverse regression.

(a)

R_z	0.25	1.0	CVR	CV
SEP	2.75	2.78	9	6
rp w.r.t.				
MLM	1.11	1.26	14	9
MLM+R	1.06	1.20	9	6

(b)

R_z	0.25	1.0	CVR	CV
SEP	12.8	5.87	17	12
rp w.r.t.				
MLM	0.63[1]	1.86	32	22
MLM+R	0.62[1]	1.82	32	22

1) one missing value because of nonconvergence of MLM for some sites

Table 8. Summary of the Imbrie and Kipp (1971) data. (SST: sea-surface temperature in °C and Salinity in 0/00, sd : sample standard deviation). For explanation of abbreviations, see Table 2 and 3.

	min	med	max
N_2	5	20	45
M_2	1	4	11

	Summer SST	Winter SST	Salinity
range	2 - 29	-1.0 - 26.5	33.5 - 37.2
sd	7.02	8.00	0.99
λ_1	0.72	0.73	0.53
λ_2	0.14	0.08	0.41
SD$_1$	3.85	4.25	2.87
SD$_2$	2.21	1.60	2.51
V^2	37.5	37.9	27.5

Table 9. Imbrie and Kipp (1971) data: apparent (RMSE) and prediction (SEP) error for winter sea-surface temperature (SST), summer SST and salinity using various methods. Superscripts denote number of components used in the estimation. (QFA: Q-mode factor analysis; CA: Correspondence Analysis; +R: followed by an inverse regression; WA_{tol}: tolerance downweighted weighted averaging calibration). For further explanation, see text.

	Summer SST		Winter SST		Salinity	
	RMSE	SEP	RMSE	SEP	RMSE	SEP
1. WA	2.02	2.21	1.97	2.14	0.57	0.60
2. WA-PLS	1.53^3	1.81	1.17^3	1.45	0.45^3	0.53
3. PLS	1.29^7	2.03	0.99^8	2.05	0.41^3	0.55
4. MLM	1.82	1.95	1.51	1.65	0.81	0.85
5. MLM+R	1.70		1.38		0.61	
Birks (unpubl.)						
Gaussian logit regr.						
6. + ML calibration	2.09		3.21		0.71	
7. + WA ,,	1.94		1.56		0.56	
8. + WA_{tol} ,,	1.80		1.25		0.53	
Imbrie & Kipp 1971						
9. QFA+R	2.55^4		2.57^4		0.57^4	
10. QFA+R (quadr.)	2.15		1.54		0.57	
Roux 1979						
11. CA+R	1.72^3		1.37^3		0.50^3	

Figure 1. L-shaped configuration of sites in the training set used in the simulations with *x* the environmental variable to be calibrated and *z* a nuisance environmental variable (L_z = 100).

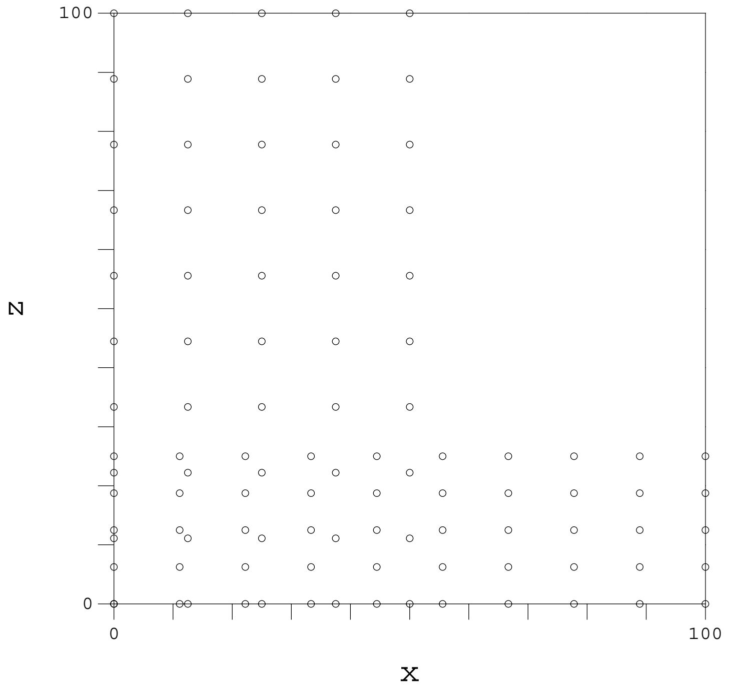


Figure 2. Abundance (percentage of the total count per site) versus the calibration variable *x* for selected species in a simulated training set for which $R_x = R_z = 1$.

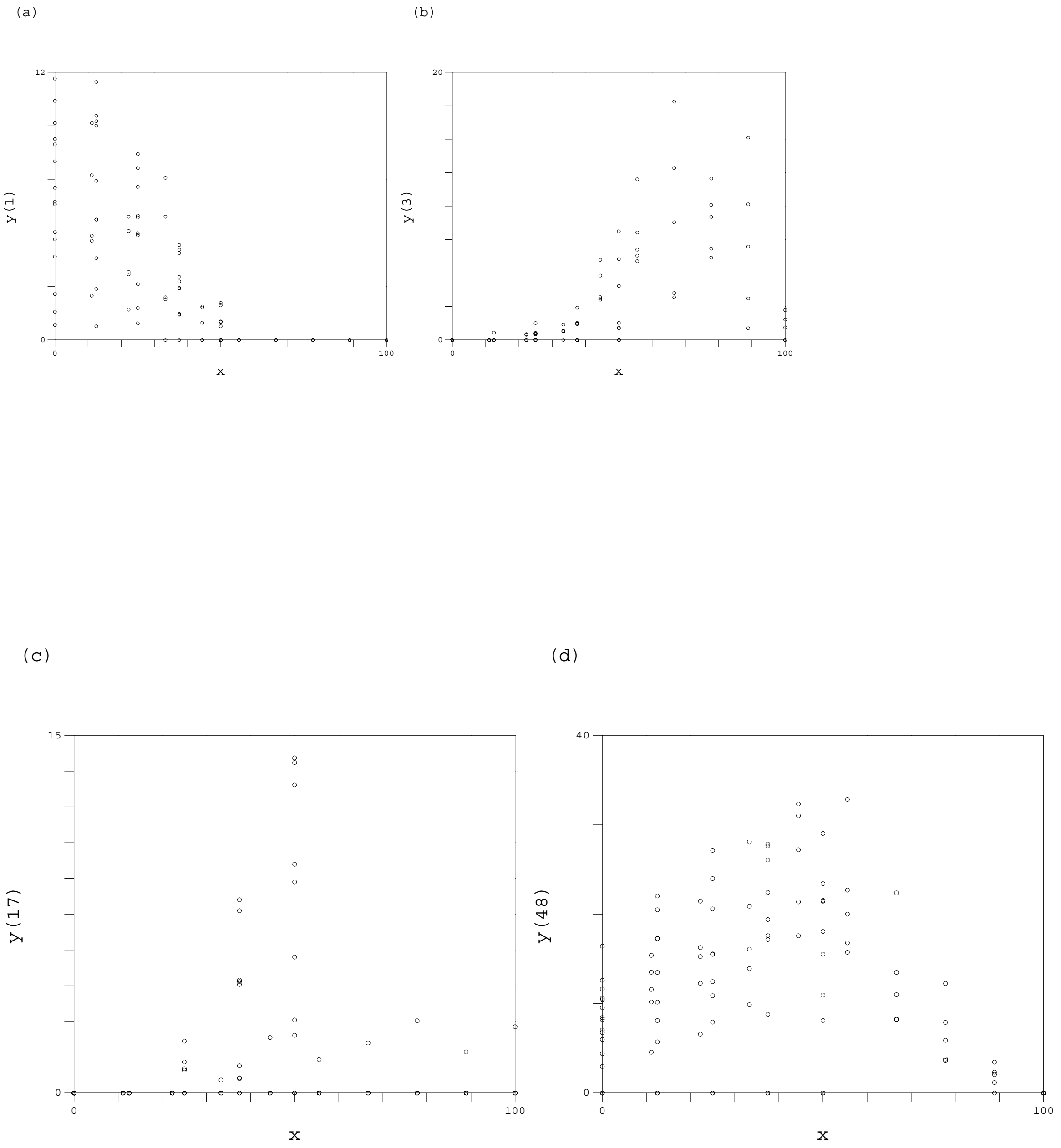


Figure 3. Residual (*r*) versus *x* after extraction of *s* WA-PLS components (*s* = 1, 2, 3 and 7). The optimal number of components is 7 (see Table 1). Data as in Figure 2.

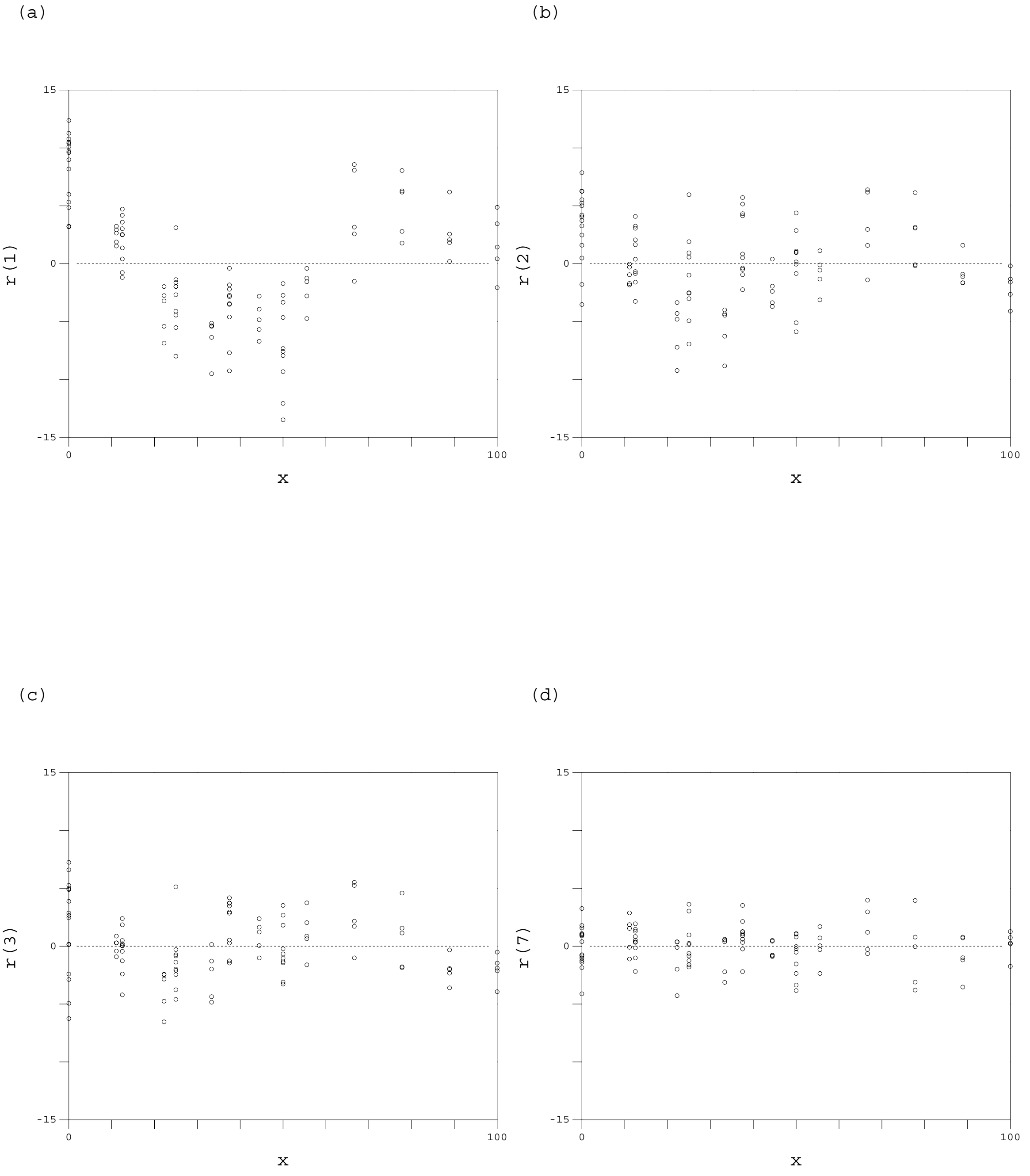


Figure 4. Predicted values (a, b) and leave-out residuals (c, d) of winter sea-surface temperature as predicted from species compositions of foraminifera after extraction of 1 (a, c) and 3 (b, d) components. The optimal number of components is 3. Data from Imbrie and Kipp (1971).

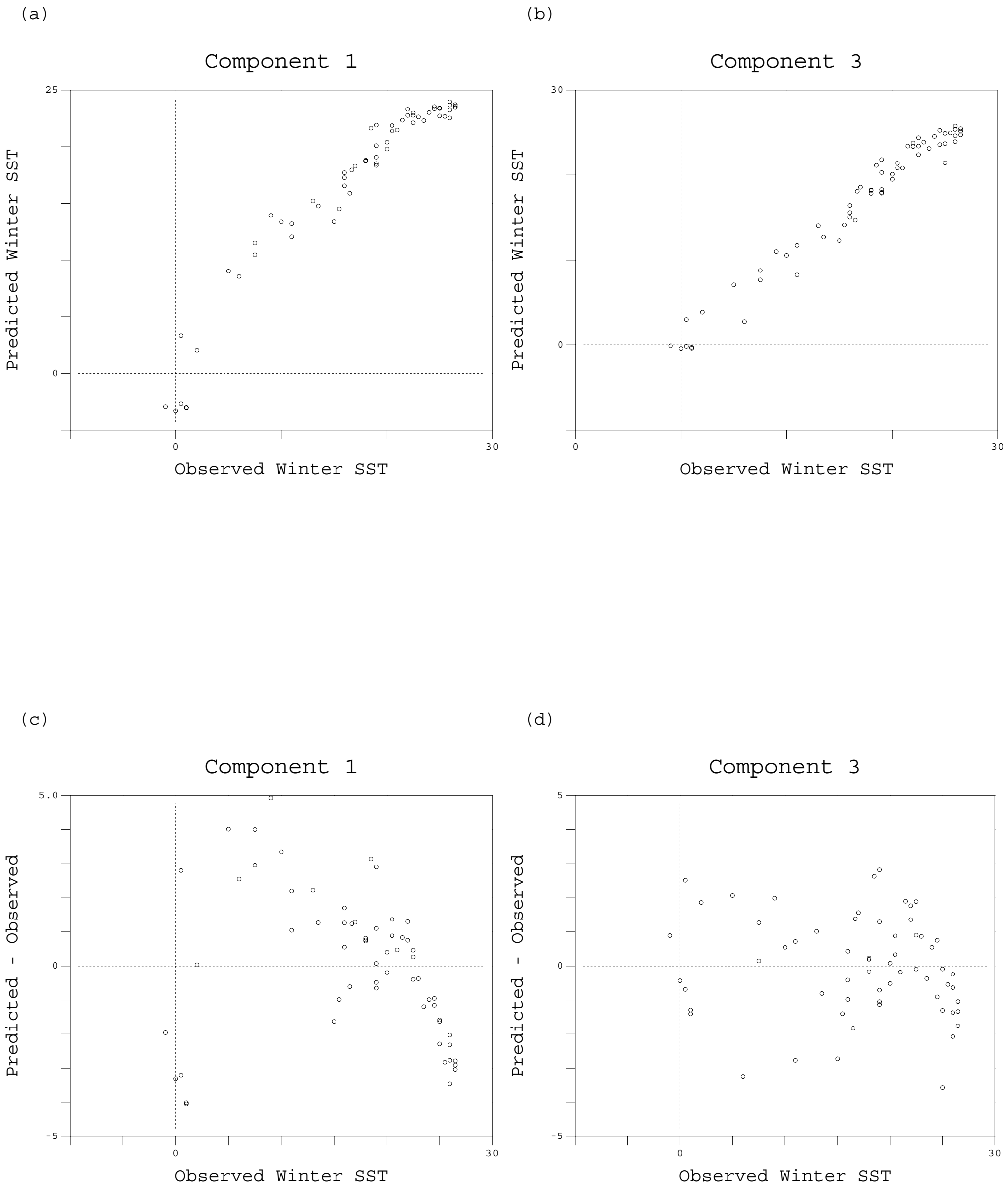
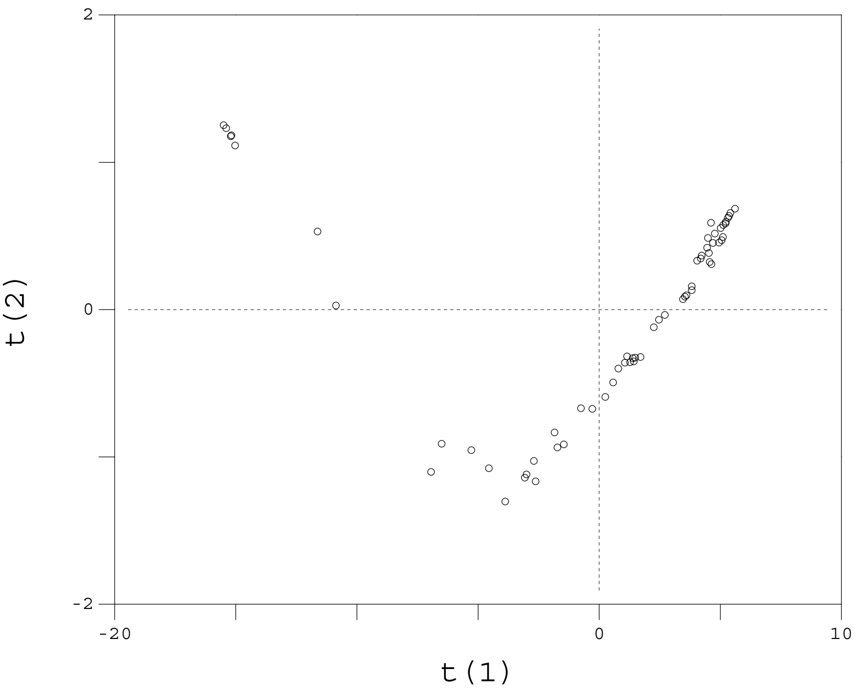


Figure 5. Plot of the first two components of WA-PLS in calibrating winter sea-surface temperature. The parabolic relation is similar to that commonly found in correspondence analysis. Data from Imbrie and Kipp (1971).


2
0
-2
t(2)
-20
0
10
t(1)