Journal of Paleolimnology **26:** 411–422, 2001.
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

A comparison between diatom-based pH inference models using Artificial Neural Networks (ANN), Weighted Averaging (WA) and Weighted Averaging Partial Least Squares (WA-PLS) regressions

Julien M.J. Racca[1], Aline Philibert[1], Robert Racca[2] & Yves T. Prairie[1]
[1]Département des sciences biologiques, Université du Québec à Montréal, Case postale 8888, succ. Centre-ville, Montréal, Canada H3C 3P8 (Email: racca.julien@courrier.uqam.ca)
[2]Département de Mathématique, Université Antilles Guyane, 97159 Pointe à Pitre Cedex, France

Received 7 December 2000; accepted 20 March 2001

Key words: inference models, artificial neural networks, WA, WA-PLS, pH, diatoms

Abstract

We explored the possibility of using artificial neural networks (ANN) to develop quantitative inference models in paleolimnology. ANNs are dynamic computer systems able to learn the relations between input and output data. We developed ANN models to infer pH from fossil diatom assemblages using a calibration data set of 76 lakes in Quebec. We evaluated the predictive power of these models in comparison with the two most commonly methods used in paleolimnology: Weighted Averaging (WA) and Weighted Averaging Partial Least Squares (WA-PLS). Results show that the relationship between species assemblages and environmental variables of interest can be modelled by a 3-layer back-propagation network, with apparent R^2 and RMSE of 0.9 and 0.24 pH units, respectively. Leave-one-out cross-validation was used to access the reliabilities of the WA, WA-PLS and ANN models. Validation results show that the ANN model ($R^2_{\text{jackknife}} = 0.63$, $\text{RMSE}_{\text{jackknife}} = 0.45$, mean bias = 0.14, maximum bias = 1.13) gives a better predictive power than the WA model ($R^2_{\text{jackknife}} = 0.56$, $\text{RMSE}_{\text{jackknife}} = 0.5$, mean bias = –0.09, maximum bias = –1.07) or WA-PLS model ($R^2_{\text{jackknife}} = 0.58$, $\text{RMSE}_{\text{jackknife}} = 0.48$, mean bias = –0.15, maximum bias = –1.08). We also evaluated whether the removal of certain taxa according to their tolerance changed the performance of the models. Overall, we found that the removal of taxa with high tolerances for pH improved the predictive power of WA-PLS models whereas the removal of low tolerance taxa lowered its performance. However, ANN models were generally much less affected by the removal of taxa of either low or high pH tolerance. Moreover, the best model was obtained by averaging the predictions of WA-PLS and ANN models. This implies that the two modelling approaches capture and extract complementary information from diatom assemblages. We suggest that future modelling efforts might achieve better results using analogous multi-model strategies.

Introduction

The first quantitative paleoenvironmental reconstruction models developed by Imbrie and Kipp (1973) were based on linear or curvilinear regression models between principal components extracted from modern species assemblages and environmental variables of interest. Since then, several other models have been proposed, two of which are now widely used in paleolimnology: Weighted Averaging regression (WA) (ter Braak & van Dam, 1989; Birks et al*.,* 1990) and Weighted Averaging-Partial Least-Squares regression (WA-PLS) (ter Braak & Juggins, 1993). These models assume a unimodal relationship between species and environmental variables. However, even though it is usual in a calibration data set that some taxa show a statistically significant unimodal or linear response to the environmental gradient of interest, other taxa may show a skewed unimodal or sigmoid increasing or decreasing response (Huisman et al., 1993; Birks, 1998). Therefore, models with a sufficient flexibility to accommodate the full range of observed responses might be more powerful than methods that assume either a unimodal response of all taxa or a linear response of all taxa (Birks, 1998).

Artificial neural networks (ANN) have the potential for modelling and incorporating such mixtures of responses. Derived from Artificial Intelligence (AI), ANNs are dynamic computer systems capable of 'learning' the relations between input and output data. They are composed of many non-linearly inter-connected simple processing units (neurons) that work in parallel. During the training process (iterative simulations), the network adapts itself from examples and the optimal relations (functions) between the input and output data are found and implemented automatically. The implemented function can then be used to predict dependent variables using only the independent ones. The main advantage in using ANN is that no *a priori* assumptions about the relation between inputs (independent variables) and output (dependent variable) are necessary. However, the drawback is that those relations learned by an ANN are hidden in its neural architecture and cannot be expressed in traditional mathematical terms.

ANNs are used in various fields including physics (Rahim et al., 1993) and medicine (Lerner et al., 1994). In ecology they are seldom used, although a few papers have shown that they can give superior results to more traditional statistical methods such as multiple regression (Brey et al., 1996; Lek et al., 1996; Moatar et al., 1999). Comparisons have also been made to paleoceanographic tools like MAT (Modern Analog Technique) (Malmgren & Nordlund, 1997), Imbrie and Kipp-type transfer functions (Malmgren & Nordlund, 1997) and SIMCA (Soft Independent Modelling of Class Analogy) (Malmgren & Nordlund, 1997).

In this paper, we propose a modelling method based on one form of neural networks: the back-propagation algorithm (Rumelhart et al., 1986). We explore the potential of this approach to modelling and inferring pH from a diatom calibration data set based on 76 lakes in Quebec. We then compare the ANN results to two other techniques commonly used in paleolimnology: WA and WA-PLS.

Methods

ANN principle

Artificial neuron

An artificial neuron is a processing element like a biological neuron (Figure 1a). It works as follows: (1) it receives input (from the original data or from the output of other neurons in the network). Each input comes via a connection which has a given strength (weight); these weights correspond to the synaptic efficiency in a biological neuron. The weighted sum of the inputs is formed to compose the activation of the neuron. (2) The activation signal is passed through an activation function (sigmoid, tan sigmoid, linear or step function) to produce the output of the neuron. The output is then duplicated as many times as needed.

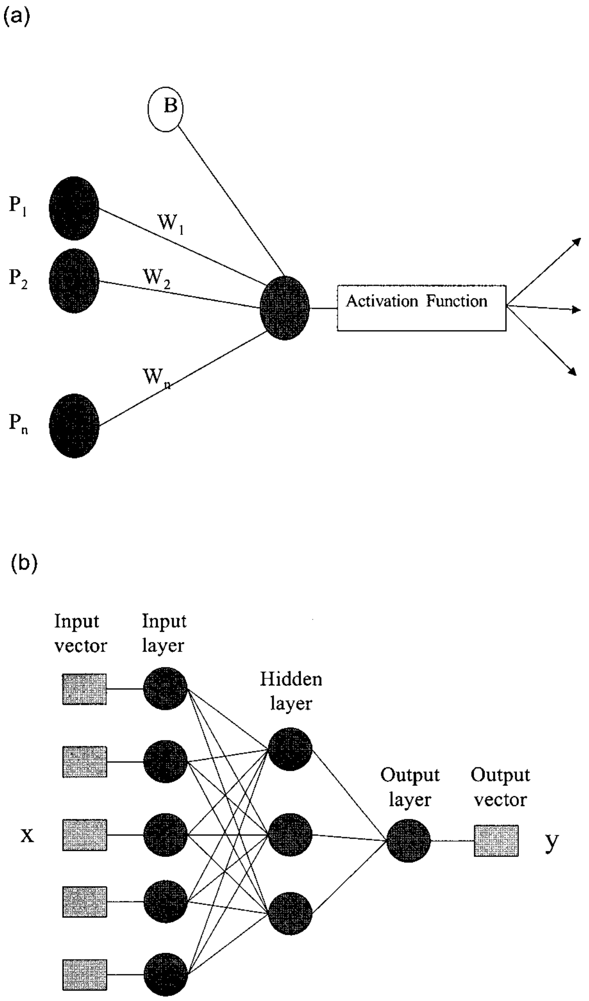

Figure 1. (a) Schematic representation of a simple processing element. The incoming signals (p) are multiplied by the weight of the connections (W) and summed. The bias (B) is then added, and the resulting sum is filtered through the activation function to produce the activity of the neuron. (b) Schematic representation of the general architecture of a 3-layer back-propagation network with five elements in the input layer, three neurons in the hidden layer, and one neuron in the output layer.

Back-propagation neural networks

In this type of network, neurons are arranged in a distinct layered topology: one input layer, one or more hidden layers and one output layer (Figure 1b). The input layer is not really neural at all: these units simply serve to introduce the value of the input variables. The hidden and output layer neurons are each connected to all of the units in the preceeding layer.

The back-propagation algorithm (descending gradient algorithm) is based on supervised learning, namely to learn, the system has to know, for each example, the output (environmental variable) associated with the input (species data). The learning phase consists of adjusting the weights of the network connections by feeding a set of input/target pattern pairs (examples) many times. The back-propagation algorithm works as follows: (1) the network is initialised by assigning a learning rate, a maximum number of iterations and random values to the synaptic weights; (2) a training pattern is fed and propagated forward through the network to compute an output value for each output unit; (3) the computed output is compared with the expected output; (4) a backward pass through the network is performed, changing the synaptic weight on the basis of the observed output errors. Steps 2 through 4 are iterated for each pattern in a training set, then the network performance is checked and a new set of training patterns is submitted to the network (i.e., a new epoch is started) if it needs further optimization. This dynamic procedure allows the difference between the predicted output and observed output to converge towards a minimal value. Details of the back-propagation algorithm are presented in Appendix 1.

Back-propagation networks are also called 'universal approximators' and, as such, they are ultimately able to learn any pattern perfectly. These networks are only really useful if they are capable, after a learning period, of generalizing. In order to generalize, a network must be able to produce the correct output data on samples not included in the learning set. A well-built neural network will, after training with a learning set, give a high proportion of correct predictions when fed a validation set. Background information on ANNs is available in various introductory textbooks such as Bishop (1995).

Data set

Data for 76 lakes distributed in two regions of Quebec (Abitibi and Réservoir Gouin) were used in this study. Lakes in Réservoir Gouin (n = 35) were sampled three times during the ice-free season in 1996 and 1997 while the Abitibi lakes (n = 41) were sampled twice between June and August 1996 (n = 20) and 1997 (n = 21) (for details, see Enache & Prairie, in press). pH values are summer averages obtained from these samples. The range, mean and median are summarised in Table 1.

Modern diatoms recovered from the surface sediments of the 76 lakes were processed, identified and counted (for methodological details, see Enache & Prairie, in press). In total, 214 fossil diatom taxa (relative abundance > 1%) were identified. Only 20% of the 214 taxa are present in at least 10 lakes while 50% are present in 3 lakes or less. An average of 18 taxa were identified per lake. Some details of the species data-set are summarised in Table 1.

Numerical methods

We determined whether to use linear- or unimodal-based regression and calibration techniques (ter Braak & Prentice, 1988; Birks, 1995) by detrended canonical correspondence analysis (DCCA; ter Braak, 1986). The gradient length of DCCA axis 1 is a measure of compositional change in the diatom data in standard deviation (S.D.) units along the pH gradient (ter Braak & Juggins, 1993; Birks, 1995). The statistical significance of the pH-diatom relationship was assessed by

Table 1. Calibration data-set characteristics

	Minimum	Maximum	Mean	Median	S.D.
Environmental variable					
pH (units)	4.16	8	6.31	6.39	0.74
Diatom					
presence (taxa/lake)	8	31	17.66	17	5.83
taxa occurrences in data-set	1	49	6.35	3	8.3

Table 2. Type of taxon responses to pH, their WA optimum, WA tolerance and number of occurences. Only taxa that occur in 5 or more lakes were considered (in total 82 taxa)

Type of response	Taxa	Optimum	Tolerance	Occurrence
Unimodal asymmetric	*Aulacoseira distans nivaloides*	5.97	0.64	11
	Aulacoseira distans tenella	6.16	0.83	33
	Cyclotella bodanica lemanica	6.62	0.52	36
	Cyclotella michiganiana	7.07	0.69	7
	Cyclotella stelligera	6.49	0.56	38
	Cymbella gaeumanii	5.62	0.62	10
	Cymbella hebridica	5.17	1.06	9
	Eunotia bilunaris	5.19	0.96	11
	Eunotia exigua	5.92	0.86	8
	Eunotia pectinalis ventralis	5.18	0.89	8
	Eunotia rhomboides	4.84	0.39	8
	Fragilaria brevistriata (form3)	6.67	0.16	9
	Fragilaria fasciculata	5.73	0.78	6
	Fragilaria virescens exigua	5.89	0.59	15
	Navicula radiosa	6.71	0.62	11
	Navicula subtillissima	5.4	0.52	12
	Neidium ampliatum	5.09	0.74	12
	Neidium iridis	5.04	1.02	9
	Tabellaria(form1)	5.96	0.7	11
Unimodal symmetric	*Amphicampa hemicyclus*	5.58	0.63	6
	Asterionella formosa	6.08	0.74	41
	Aulacoseira alpigena	6.47	0.41	17
	Aulacoseira italica subarctica	6.51	0.4	38
	Brachysira brebissonii (form1)	6.07	0.51	33
	Cyclotella comensis	6.39	0.4	8
	Cyclotella ocellata	6.73	0.32	11
	Cyclotella pseudostelligera (form3)	7.06	0.45	5
	Cymbella gracillis	6.17	0.53	10
	Eunotia bidentula	4.66	0.35	7
	Eunotia incisa	6.13	0.48	20
	Eunotia pectinalis	6.38	0.35	9
	Fragilaria intermedia	7.24	0.36	5
	Fragilaria pinnata pinata	7.19	0.46	10
	Frustulia rhomboides var. saxonica	6.28	0.35	24
	Navicula mediocris	6.11	0.41	5
	Tabellaria flocculosa (form3)	6.29	0.54	49
	Tabellaria flocculosa (form4)	4.97	0.3	7
	Tabellaria fenestrata	6.46	0.32	22
	Tabellaria quadrisepta	6.29	0.33	9
	Tabellaria ventricosa	6.37	0.15	10
Sigmoid increasing	*Achnanthes minutissima minutissima*	6.71	0.99	42
	Amphora ovalis	6.79	0.42	7
	Brachysira vitrea	6.78	0.95	14
	Eunotia tenella	6.74	0.6	16
	Fragilaria brevistriata (form1)	6.73	0.78	7
	Fragilaria construens binodis	6.88	0.86	17
	Fragilaria construens venter	7.04	1.06	14
	Fragilaria lata	6.65	1.16	10
	Gomphonema gracile	6.77	0.61	7
	Navicula pupula pupula	6.73	0.61	10
Sigmoid decreasing	*Actinella punctata*	5.2	0.92	6
	Aulacoseira nygaardii	5.32	0.98	6
	Aulacoseira perglabra	5.32	1.97	16

Table 2. Continued

Type of response	Taxa	Optimum	Tolerance	Occurrence
	Eunotia faba	5.85	0.94	7
	Frustulia rhomboides	5.47	1.02	25
	Pinnularia braunii	5.3	1.34	15
	Pinnularia gibba	5.81	1.02	14
	Pinnularia microstauron (form1)	5.84	1.15	6
	Pinnularia microstauron (form2)	5.63	0.76	11
	Stauroneis agrestis	5.7	1.03	6
	Tabellaria ventricosa	5.78	0.93	24
No relation	*Aulacoseira ambigua*	6.47	1.07	11
	Aulacoseira distans distans	6.53	0.92	22
	Aulacoseira distans humilis	6.34	0.65	14
	Aulacoseira lirata	6.2	0.67	25
	Brachysira brebissonii (form3)	5.88	0.47	5
	Cyclotella meneghiniana	6.69	0.56	5
	Cyclotella pseudostelligera (form1)	6.32	0.75	8
	Cymbella incisa	6.38	0.44	5
	Cymbella microcephala	6.53	1.46	8
	Cymbella silesiaca	6.3	0.46	9
	Eunotia bilunaris mucophyla	6.01	0.69	19
	Eunotia paludosa	6.43	0.73	8
	Eunotia polyglyphis	6.67	1.19	6
	Eunotia subarcuatoides	5.99	0.65	6
	Fragilaria nanana	6.471	0.81	5
	Navicula subatomoides	6.5	0.84	9
	Navicula laevissima	6.45	0.9	9
	Nitzschia fonticola	6.35	1.18	8
	Nitzischia minutula	6.46	0.68	12
	Pinnularia maior	6.49	0.58	8
	Stauroneis phoenicenteron	6.04	1.13	10

Monte Carlo permutation tests involving 99 unrestricted permutations (ter Braak, 1990). DCCA was implemented using CANOCO version 3.12 (ter Braak, 1988, 1990).

We also evaluated the relationship to pH of all taxa occurring in a least 5 lakes (82 taxa). The statistical relationship was assessed using a hierarchical set of taxon response models (Huissman et al., 1993). This hierarchical set of response models consists of a skewed asymmetric unimodal response model, a symmetric Gaussian unimodal response model (ter Braak & Looman, 1986; Birks et al., 1990), a monotonically increasing or decreasing sigmoidal response model (Birks et al., 1990), and a null model with no relationship to pH (Birks et al., 1990). The taxon response modelling was done using the program HOF (J. Oksanen, unpublished program). Taxa with statistically significant fits to pH under the different types of response models are given in Table 2.

WA and WA-PLS models

As the data-set had a gradient greater than 2.5 S.D units along its pH gradient, the unimodal techniques of Weighted Averaging (WA) and Weighted Averaging Partial Least Square (WA-PLS) were used to develop pH diatom-based inference models. In WA regression, the optimum for each taxon is estimated from the training set based on the abundance of diatoms in the surficial sediment and the measured environmental variables (Birks et al., 1990). The regression step then allows inference of the environmental conditions from the diatom composition (Birks et al., 1990). WA-PLS is an extension of simple WA (ter Braak & van Dam, 1989; Birks et al., 1990) in which successive components are extracted from the training set. WA and WA-PLS were carried out using a SAS/IML implementation of the algorithm (Y.T. Prairie, unpublished program). In this program, the final number of components retained in the model corresponds to the minimum number of components whose jackknifed RMSE is not significantly (as assessed by an F ratio on the MSEs) higher than the model with the minimum $\mathrm{RMSE}_{\mathrm{jackknife}}$ (van der Voet, 1994).

ANN models

To compare with the WA and WA-PLS models, we used back-propagation neural networks with 3 layers. Each unit of the first layer (input) is an identity neuron. Their activity represents the values of the relative abundance (%) of the diatom species in a sample. The output layer is composed of a unique unit (linear activation function) representing pH. The hidden layer contains units with symetrical-sigmoid activation functions. The number of units in this layer depends on the complexity of the problem. In order to find the best possible network, we tried various numbers of hidden layer units (2, 3, 5, and 10). So-called bias neurons, connected to each neuron on the hidden layer and the output layer were also used. This type of neuron is similar to a constant in a multiple regression. The pH neural network models were built using YANNS (Yet Another Neural Network Simulator) (Boné et al., 1998).

Models validation

WA and WA-PLS

The predictive ability of these diatom-based calibration models was assessed by the coefficient of determination between the measured and the diatom-inferred values (R^2) and the apparent root mean square error (RMSE). However, $R^2_{\mathrm{jackknife}}$ and $\mathrm{RMSE}_{\mathrm{jackknife}}$ were also computed as they are more realistic measures of predictive power than the apparent statistics (Birks, 1998). Jackknifing consists of an iterative re-sampling technique involving a new training set of n-1 lakes from the original calibration set and its application to the one excluded sample. Further aspects of the model performance are the average bias and the maximum bias in the residuals for the test set (ter Braak & Juggins., 1993). For estimation of maximum bias, the sampling interval was subdivided into 10 equal intervals, the bias per interval calculated and the maximum of the 10 values calculated (ter Braak & Juggins., 1993).

ANN

Given the relatively low number of lakes (76) in this study, usual cross-validation methods (K-fold cross-validation or Hold-out procedure (German et al., 1992)) were not appropriate. These methods consist of randomly dividing the calibration set into two subsets (learning and validation 1:1, 2:1 or 3:1, etc..) and are not well-suited to short and large data sets typical of paleolimnological applications. This is because the training set still has to be large enough to be representative and the validation set has to be large enough to allow a robust validation of the network. We therefore used the same validation method as used in WA and WA-PLS models, namely leave-one-out cross-validation or jackknifing cross-validation (Efron, 1983; Kohavi, 1995). In this case, each lake in the calibration data-set (species/environmental variables couplets) is successively used in the validation. The jackknifing technique consists of building a complete set of networks (each with n-1 training examples and 1 validation case) and attempts to find, for the entire set of networks, a common number of iterations for an optimal generalization. This generalization is expressed as the average $\mathrm{RMSE}_{\mathrm{jackknife.}}$ The number of iterations used in the construction of the final network uses the early stopping method, that is when the average error in the validation set is minimal. This avoids overfitting (overtraining). Apparent RMSE is given by the average error in the learning set when training is stopped. $\mathrm{RMSE}_{\mathrm{jackknife}}$ is given by the average error in the validation set. We also evaluated the average and the maximum bias of the residuals as previously described.

Results and discussion

Data-set characteristics

The diatom data-set is large (214 taxa × 76 lakes), sparse (8–31, mean = 18 taxa per sample, 91% of zero values) and can therefore be characterized as noisy. pH values are normally distributed in the data set (Shapiro-Wilk W statistic = 0.96, prob (W) = 0.06) and the range of the observed values is large (4.16–8.05 mean = 6.31). Fifty percent of samples in the data-set have a pH between 6 and 6.74. Only 5 lakes have a pH < 5 and 10 lakes a pH > 7. Table 1 shows the details of the calibration data-set characteristics.

Monte Carlo permutation tests of DCCA axis 1 constrained to pH show that the data-set has a statistically significant (F ratio = 3.45, p value = 0.01) relationship to pH. pH explains 4.5% of the variance in the data-set and the gradient length is 3.55 S.D.

Table 2 shows the type of response of all taxa present in 5 lakes or more (40% of total species). Seventy five percent of these taxa have statistically significant responses to pH; 50% have a symmetric Gaussian or asymmetric unimodal responses while 25% have an increasing or decreasing sigmoidal response. Given the long gradient length (3.55 S.D., pH range of 3.86 units), it is surprising that only 50% of taxa have a unimodal response to pH.

ANN pH-inference models

Optimal performances were obtained with 3 units in the hidden layer (214/3/1) (Table 3). RMSE decreases exponentially as the number of training epochs (iterations) increases while $\text{RMSE}_{\text{jackknife}}$ decreases and then increases (Figure 2). The best pH model was obtained after 700 epochs (Figure 3). RMSE and $\text{RMSE}_{\text{jackknife}}$ were 0.24 and 0.45, respectively. The apparent coefficient of determination (R^2) was 0.9 while $R^2_{\text{jackknife}}$ was 0.63. Figures 3a and 3b illustrate the fit between the estimated or jackknifed-predicted and observed pH when RMSE and $\text{RMSE}_{\text{jackknife}}$ are minimal. Figures 3c and 3d illustrate the distribution and the homogeneity of residuals. Residuals are normally distributed (Shapiro-Wilk W test = 0.98 prob W ≈ 0.90) and the average bias and maximum bias are, respectively, 0.14 and 1.13 pH units. We did not observe any systematic trend in the residuals (Figure 3d).

WA and WA-PLS pH-inference models

WA pH-inference models gave an apparent and jackknifed RMSE of, respectively, 0.33 and 0.5 pH units, with corresponding apparent R^2 and $R^2_{\text{jackknife}}$ of 0.81 and 0.56. Other summary statistics are given in Table 3. Plots of observed pH against estimated (apparent) (Figure 4a), or jackknifed-predicted pH (Figure 4b), and their corresponding residual plots (Figures 4c & 4d) show that there is a systematic bias in the predictive model, with a tendency for predicted values to be over-estimated at low pH and under-estimated at high pH values. The range of predictions is almost 2 times smaller (5.17–7.26, 2.09 units) than the observed range (4.16–8.0, 3.84 units). This bias disappeared after inverse deshrinking (Birks et al., 1990).

Table 3. Descriptive ANN, WA and WA-PLS pH-models summaries

	ANN	WA	WA-PLS
Number of components		1	3
Number of hidden units (ANN)	3		
Number of samples used	76	76	76
Number of taxa used	214	214	214
Number of iterations (ANN)	700		
$\text{RMSE}_{\text{apparent}}$ (pH units)	0.24	0.33	0.23
$\text{RMSE}_{\text{jackkinfe}}$ (pH units)	0.45	0.5	0.48
R^2_{apparent}	0.9	0.81	0.9
$R^2_{\text{jackknife}}$	0.63	0.56	0.58
Mean bias	0.14	–0.09	–0.15
Max bias	1.13	–1.07	–1.08

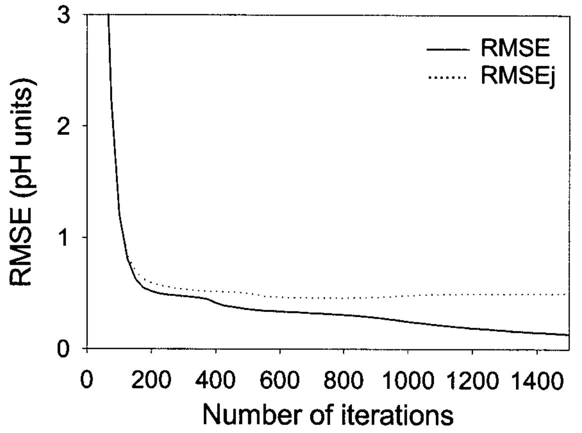

Figure 2. Changes in root-mean-square error of estimation (apparent RMSE) and root-mean-square error of prediction ($\text{RMSE}_{\text{jackknife}}$) for pH with increasing number of iterations in 3-layer back-propagation neural network with 3 units in the hidden layer. The netwoks were trained over 60 intervals of 25 iterations each (in total 1,500 iterations).

Three-component WA-PLS pH-inference models gave slightly better results than WA: 0.23 pH units for RMSE, 0.48 for $\text{RMSE}_{\text{jackknife}}$ while R^2 and $R^2_{\text{jackknife}}$ are 0.90 and 0.58, respectively (Table 3). Figures 5a and 5b show the fit between the observed and estimated pH values and the observed and jackknifed-predicted pH values. Mean bias and maximum bias are, in this case, –0.15 and –1.08 (Table 3). The WA-PLS model out-performs the WA model mainly because the high pH values are not underestimated (Figure 5d). In this case, the predicted range (4.64–8.19, 3.55 units) is almost as large as the observed range. Because the number of components of the final WA-PLS model is chosen on the basis of the decreasing RMSE with successive components before deshrinking, it is quite conceivable that the secondary components of any WA-PLS models serve mostly as a deshrinking component. Contrary to ANN models, the residuals from the WA and WA-PLS models were not normally distributed (Figures 4c & 5c).

Taxon inclusion in the models

How many taxa to include in an inference model has already been addressed to some extent by Birks (1994)

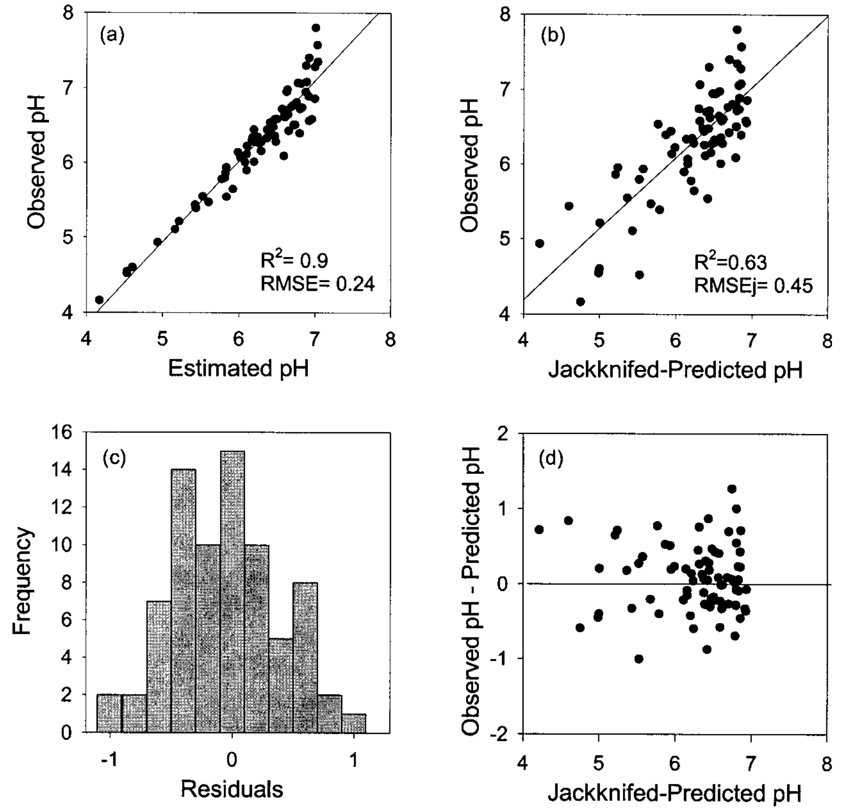

Figure 3. (a) Plot of observed pH values for the 76 lakes within the calibration set against estimated pH values. The estimated values are based on the training set of a 3-layer back-propagation network with three units in the hidden layer. (b) Plot of observed pH values against jackknifed-predicted pH values. The jackknifed-predicted values are based on the validation set (leave-one-out). For (a) and (b), the fitted line is based on Model I regression. (c) Diagram of the distribution of the residuals (from b). (d) Plot of residuals (from b) against predicted pH.

and Wilson et al. (1996). Inclusion criteria can be based either on a minimum abundance limit (i.e., 1% relative abundance) or based on the occurrence of the taxa in a minimum number of samples. Here we present an analysis based on the taxon's tolerances calculated prior to the development of the models. A series of numerical experiments was done to see if deletion based on tolerance resulted in any change in WA-PLS and ANN performances. For this purpose, only taxa that occur in 5 or more lakes were considered.

As a high tolerance is indicative of ubiquitous taxa and therefore unlikely to be affected by environmental changes, we successively deleted from the data matrix taxa with the highest tolerances (0.9, 0.8, . . ., 0.5). For the ANN models, this procedure led to a slight reduction in performance (as revealed by the $R^2_{\text{jackknife}}$, Figure 6a) but, overall, the ANN performance remained very stable. This was not the case for WA-PLS models, where the performance steadily improved as taxa with high tolerances were removed. Ultimately, the best fit was achieved when taxa with tolerances above 0.6 pH units were omitted ($\text{RMSE}_{\text{jackknife}}$ = 0.40 pH units and $R^2_{\text{jackknife}}$ = 0.72). It is interesting to note that the taxa with a high tolerance are essentially the taxa with a sigmoidal response curve as opposed to a unimodal response, as WA-PLS assumes.

We then evaluated in the same general manner whether the deletion of taxa with narrow tolerances would affect the predictive capacities of the WA-PLS and ANN models. Although not biologically surprising, our results show that the removal of non-ubiquitous taxa (e.g., taxa with a unimodal response curve) greatly affected the predictive power of WA-PLS models (Figure 6b). However, this procedure hardly affected the performance of the ANN model (Figure 6b). These results demonstrate that, unlike WA-PLS inference models, it is possible for ANN to infer pH from taxa with a high tolerance. Clearly, the taxa information used by the ANN is substantially different, both numerically and conceptually, from that used by WA-PLS.

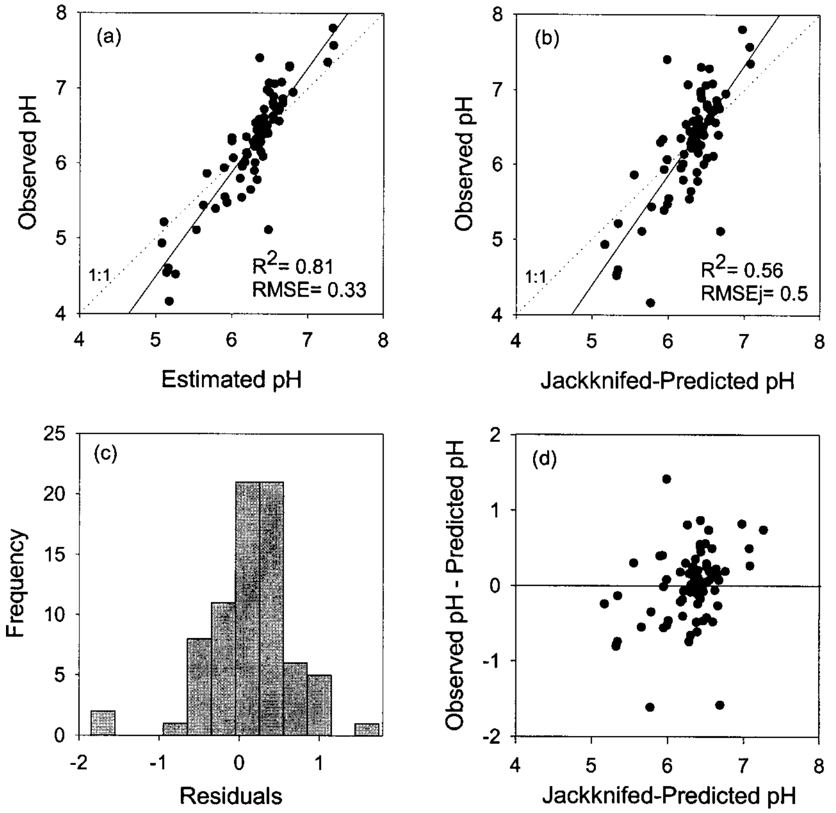

Figure 4. (a) Plot of observed pH values for the 76 lakes within the calibration set against estimated pH values. The estimated values are based on weighted averaging (WA) regression. (b) Plot of observed pH values against jackknifed-predicted pH values. The jackknifed-predicted values are based on a leave-one-out. For (a) and (b), the fitted line is based on Model I regression. (c) Diagram of the distribution of the residuals (from b). (d) Plot of residuals (from b) against predicted pH.

Comparison of the models

It is important to note that the WA-PLS and ANN models differ in two major ways. First, contrary to WA-PLS models, ANNs do not require that taxa show a unimodal relation to pH to obtain good results. In fact, the inclusion or exclusion of taxa depending on their tolerance indicates that the WA-PLS performance depends largely on the percentage of low-tolerance taxa within the calibration set. Second, even if the results show that the predictive abilities of WA-PLS and ANN models are relatively similar in global terms such as $RMSE_{jackknife}$ or $R^2_{jackknife}$ when all taxa are used, we observed that the predictions for a given lake can be very different between the two predictive models. Figure 7 illustrates the similarities and differences between the two models as a plot of ANN predictions vs. WA-PLS predictions. Predicted pH can differ by nearly one pH unit for some lakes depending on the model used. The average absolute difference between the models' predictions was 0.30 pH unit.

This reinforces the notion that the two types of models are not only different mathematically, they are also different in the taxon information used in their prediction. They should therefore be viewed as complementary models. This is further demonstrated by the fact that the model based on the average prediction of the two models is better ($R^2_{jackknife}$ = 0.74, $RMSE_{jackknife}$ = 0.38, mean bias = 0.005, and maximum bias = –0.71) than either of the models alone (Figure 8). When we combined the predictions obtained from the best WA-PLS model (with wide-tolerance taxa removed) with the best ANN model, the predictions were again improved. Clearly, the information extracted from the diatom assemblage data is not implemented in the same way by the WA-PLS and ANN algorithms. It suggests that each model is capable of capturing a part, but not all, of the underlying complex relationships between diatom assemblages and pH. If this is the case, the development of dual models, based on the average re-

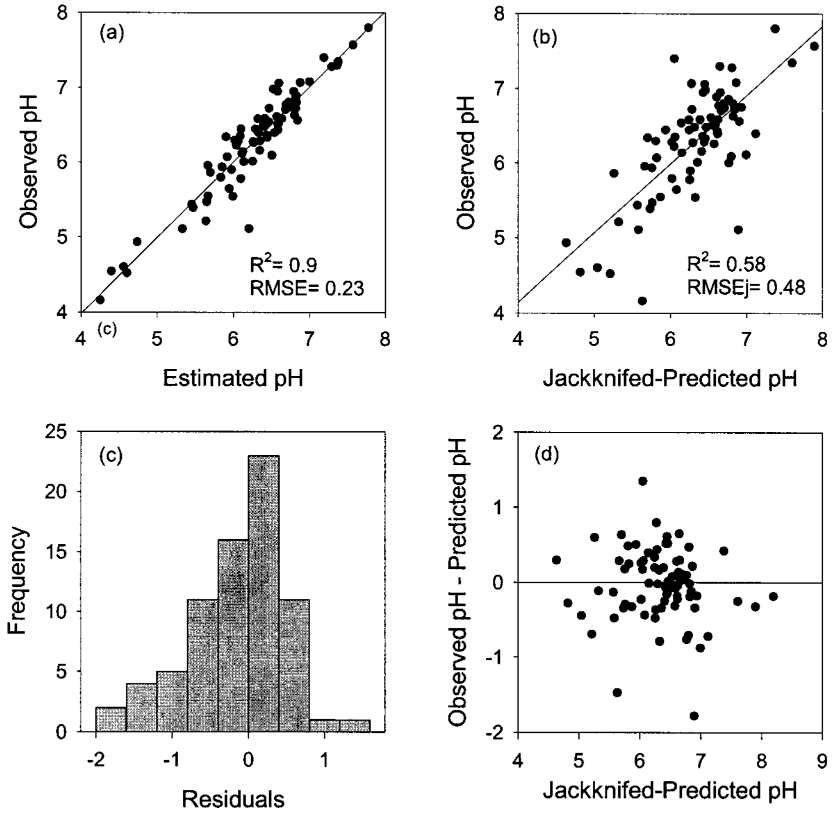

Figure 5. (a) Plot of observed pH values for the 76 lakes within the calibration set against estimated pH values. The estimated values are based on a three-component weighted averaging partial least square (WA-PLS) regression. (b) Plot of observed pH values against jacknifed-predicted pH values. The jacknifed-predicted values are based on a leave-one-out. For (a) and (b), the fitted line is based on Model I regression. (c) Diagram of the distribution of the residuals (from b). (d) Plot of residuals (from b) against predicted pH.

sults of both WA-PLS and ANN, may become a tedious but necessary procedure to obtain more reliable and robust reconstructions. The next step will be to compare results of this multi-model approach to those obtained from the increasingly popular, but even more tedious, multi-proxy models (Lotter et al., 1998). We

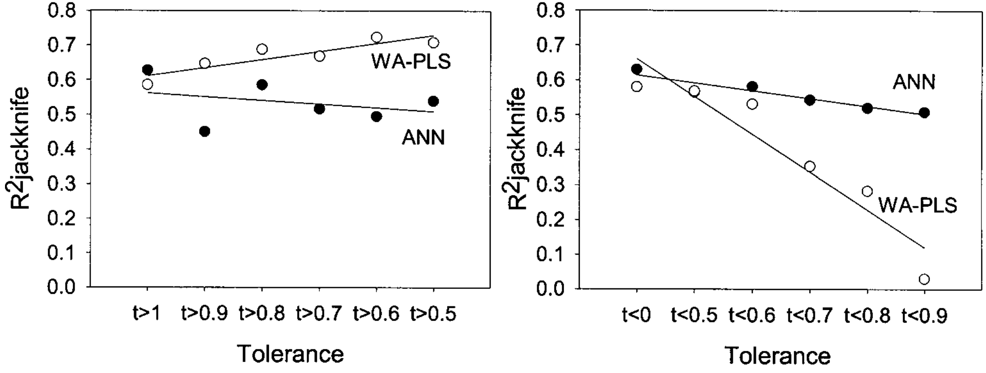

Figure 6. Plots illustrating the changes of R$^2_{\text{jackknife}}$ (a) when taxa with high tolerance are progressively removed (0.9, 0.8, . . ., 0.5 pH units) and (b) when taxa with low tolerance (0.5–0.9) are progressively removed.

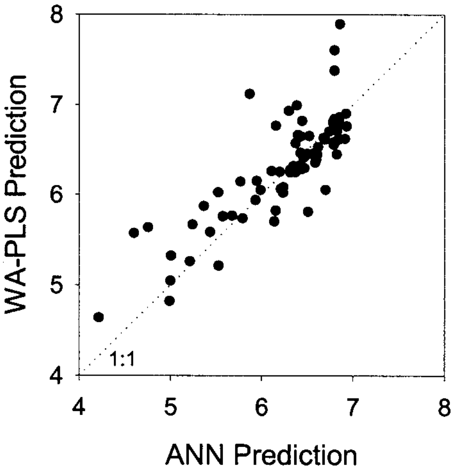

Figure 7. Plots illustrating the differences between ANN and WA-PLS predictions for each lake in the calibration data-set.

are presently attempting to generalize our results to the prediction of variables other than pH.

Conclusion

In this paper, we introduced the application of ANNs to paleolimnological pH reconstruction based on diatoms. Our comparison of the relative performance of WA-PLS and a three-layer back-propagation network models on a pH-diatom data-set from 76 lakes showed that Artificial Neural Networks can provide reliable paleolimnological inference models and that their predictive power is similar to that obtained from WA-PLS. However, they also differed on a number of points. WA-PLS is much more sensitive to taxon deletion based on their tolerance levels than ANNs. The two types of models appear to differ in the way information is extracted from the biological data and, as a result, they are complementary. Dual models produced the best predictive models.

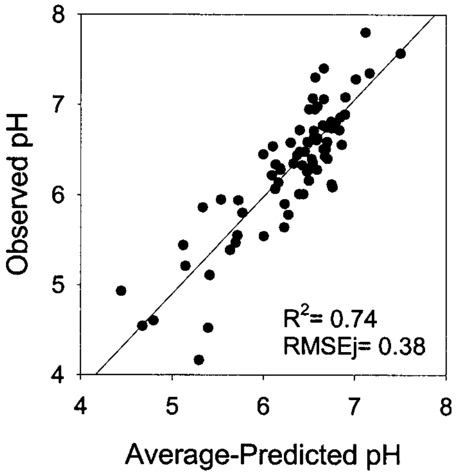

Figure 8. Plot of observed pH values for the 76 lakes within the calibration set against average-predicted pH values. The predicted values are based on the average prediction of the two models (WA-PLS and the 3-layered back-propagation networks) when all taxa are used. The fitted line is based on Model I regression.

Acknowledgements

This work is a contribution to the GREAU (Groupe de Recherche en Écologie Aquatique de l'UQAM) and GRIL. It was supported by a grant from NSERC to Y.T.P. We would like to thank John Birks for his insightful comments on the manuscript. Richard Carignan generously made available the chemical data from the Haute-Mauricie lakes. Suggestions from Matthew Wild were also greatly appreciated.

Appendix

A brief algorithm of back-propagation in neural networks. Adapted from Lek et al. (1996).

1. Initialize the number of hidden nodes.
2. Initialize the maximum number of iterations and the learning rate (η). Set all weights to small random numbers.
3. For each training vector (input $X_p = (x_1, x_2, \ldots x_n)$, output Y) repeat steps 4–7.
4. Present the input vector to the input nodes and the output to the output node.
5. Calculate the input to the hidden (h) nodes:

$$a_j^h = \sum_{i=1}^{n} W_{ij}^h \cdot x_i \qquad (1)$$

with a_j: activation of the jth downstream neuron, x_i: value at the outlet of the ith neuron of the first layer (relative abundance of taxon i), W_{ij}: weight of the connection between the ith neuron of the first layer and jth neuron of the hidden layer

Calculate the output from the hidden nodes:

$$x_j^h = f(a_j^h) = \frac{\exp(a_j^h) - 1}{\exp(a_j^h) + 1} \qquad (2)$$

Calculate the input to the output (k) node:

$$a_k = \sum_{j=1}^{L} W_{jk} \cdot x_j^h \text{ (with } L\text{: number of hidden nodes)} \qquad (3)$$

and the corresponding output:

$$\hat{Y}_k = f(a_k) = a_k \text{ (notice that in our case } k\text{=1 and } \hat{Y}_k = \hat{Y}\text{)} \quad (4)$$

6. Calculate the error term for the output node:

$$\delta_k = (Y - \hat{Y}) \quad (5)$$

and for the hidden nodes:

$$\delta_j^h = f'(a_j^h) \sum_k \delta_k W_{jk} \quad (6)$$

7. Update weights on the output layer:

$$W_{jk}(t+1) = W_{jk}(t) + \eta \delta_k x_j^h \quad (7)$$

and the hidden layer:

$$W_{ij}(t+1) = W_{ij}(t) + \eta \delta_j^h x_i \quad (8)$$

While network errors are larger than some preferred limit or the number of iterations is smaller than the maximum number of iterations, repeat steps 4–7.

References

Birks, H. J. B., 1994. The importance of pollen and diatom taxonomic precision in quantitative palaeoenvironmental reconstructions. Rev. Paleobot. Palynol. 83: 107–117.

Birks, H. J. B., 1995. Quantitative palaeoenvironmental reconstructions. In Maddy D. & J. S. Brew (eds), Statistical Modelling of Quaternary Science Data, Technical Guide 5. Quaternary Research Association, Cambridge, 161–254.

Birks, H. J. B., 1998. Numerical tools in paleolimnology: progress potentialities and problems. J. Paleolim. 20: 307–332.

Birks, H. J. B., J. M. Line, S. Juggins, A. C. Stevenson & C. J. F. ter Braak, 1990. Diatoms and pH reconstructions. Phil. Trans. r. Soc. Lond. B 327: 263–278.

Bishop, C. M., 1995. Neural networks for pattern recognition. Oxford Clarendon Press, Oxford, 482 pp.

Boné, R., M. Crucianu & J.-P. Asselin de Beauville, 1998. Yet Another Neural Network Simulator, Proceedings of the conference NEURal networks and their Applications (NEURAP'98). Marseilles, France, 421–424.

Brey, T., A. Jarre-Teichmann & O. Borlich, 1996. Artificial neural network versus multiple linear regression: Predicting P/B ratios from empirical data. Mar. Ecol. Prog. Ser. 140: 251–256.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross validation. J. Am. Stat. Assoc. 78: 316–330.

Enache, M. & Y. T. Prairie, in press. Diatom assemblages and their relationship to environmental variables in lakes from Abitibi (Québec, Canada). J. Paleolim. (in press).

German, S., E. Bienenstock & R. Doursat, 1992. Neural networks and the bias/variance dilemma. Neural Comp. 4: 1–58.

Huisman, J., H. Olff & L. F. M. Fresco, 1993. A hierarchical set of models for species response analysis. J. Vegetation Sci. 4: 37–46.

Imbrie, J. & Kipp, N. G., 1973. A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Carribean core. In Turekian, K. K. (ed.), The Late Cenozoic Glacial Ages. Yale University Press, New Haven and London, 71–181.

Kohavi, R., 1995. A study of cross validation and bootstrap for estimation and model selection. Proc. 14th Int. Joint Conf. On Artificial Intelligence. Morgan Kaufmann Publishers, 1137–1143.

Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga & S. Aulagnier, 1996. Application of neural networks to modelling non linear relationships in ecology. Ecol. Model. 90: 39–52.

Lerner, B., M. Levinstein, B. Rosenberg, H. Guterman, I. Dinstein & Y. Romem, 1994. Feature selection and chromosomes classification using a multilayer perceptron neural network., IEEE Int. Confer. on Neural Networks, Orlando (Florida), pp. 3540–3545.

Lotter, A. F., H. J. B. Birks, W. Hofmann & A. Marchetto, 1998. Modern diatom, cladocera, chironomid, and chrysophyte cyst assemblages as quantitative indicators for reconstruction of past environmental conditions in the Alps. II. Nutrients. J. Paleolim. 19: 443–463.

Malmgren, G. & U. Nordlund, 1997. Application of artificial neural networks to paleoceanographic data. Palaeogeogr. Palaeoclim. Palaeoecol. 136: 359–373.

Moatar, F., F. Fessant & A. Poirel, 1999. pH modelling by neural networks. Application of control and validation data series in the Middle Loire river. Ecol. Modelling 120: 141–156.

Rahim, M. G., C. C. Goodyear, W. B. Kleijn, J. Schroeter & M. M. Sondhi, 1993. On the use of neural networks in articulatory speech synthesis. J. Acoustical Soc. Am. 93: 1109–1121.

Rumelhart, D. E., G. E. Hinton & R. J. Williams, 1986. Learning representation by back-propagating errors. Nature 323: 533–536.

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., 1988. CANOCO: a FORTRAN program for canonical community.ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis, and redundancy analysis (version 2.1). Report LWA-88-02, Agricultural Mathematics Group, Wageningen, 95 pp.

ter Braak, C. J. F., 1990. Update Notes; CANOCO-version 3.10. Agricultural Mathematics group, Wageningen, 35 pp.

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.

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 269/270: 485–502.

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. & I. C. Prentice, 1988. A theory of gradient analysis. Adv. Ecol. Res. 18: 271–317.

Van der Voet, H., 1994. Comparing the predictive accuracy of models using a simple randomization test. Chemometr. Intel. Lab. Syst. 25: 313–323.

Wilson, S. E., B. F. Cumming & J. P. Smol, 1996. Assessing the reliability of salinity inference models from diatoms assemblages: an examination of 219-lake data set from western North America. Can. J. Fish. aquat. Sci. 53: 1580–1594.