Journal of Paleolimnology **29:** 123–133, 2003.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Separating wheat from chaff: Diatom taxon selection using an artificial neural network pruning algorithm

Julien M.J. Racca[1,2,*], Matthew Wild[1], H.J.B. Birks[3,4] and Yves T. Prairie[1]
[1]Département des Sciences Biologiques, Université du Québec à Montréal, Case Postale 8888 succ. Centre-Ville, Montréal (QC) H3C 3P8, Canada; [2]Current address: Paleolimnology-Paleoecology Laboratory, Centre d'études nordiques & Département de Géographie, Université Laval, Québec (QC) G1K 7P4, Canada; [3]Botanical Institute, University of Bergen, Allégaten 41, N-5007 Bergen, Norway; [4]Environmental Change Research Centre, University College London, 26 Bedford Way, London WC1H OAP, UK; [*]Author for correspondence (e-mail: racca.julien@courrier.uqam.ca)

Received 15 March 2002; accepted in revised form 1 August 2002

Key words: Artificial Neural Networks, Diatom, Model robustness, Pruning, Taxa contribution

Abstract

This study addresses the question of what diatom taxa to include in a modern calibration set based on their relative contribution in a palaeolimnological calibration model. Using a pruning algorithm for Artificial Neural Networks (ANNs) which determines the functionality of individual taxa in terms of model performance, we pruned the Surface Water Acidification Project (SWAP) pH-diatom data-set until the predictive performance of the pruned set (as assessed by a jackknifing procedure) was statistically different from the initial full-set. Our results, based on the validation at each 5% data-set reduction, show that (i) 85% of the taxa can be removed without any effect on the pH model calibration performance, and (ii) that the complexity and the dimensionality reduction of the model by the removal of these non-essential or redundant taxa greatly improve the robustness of the calibration. A comparison between the commonly used ''marginal'' criteria for inclusion (species tolerance and Hill's N2) and our functionality criterion shows that the importance of each taxon in an ANN palaeolimnological model calibration does not appear to depend on these marginal characteristics.

Introduction

Several types of algorithm have been proposed to develop quantitative inference models in palaeolimnology (Birks 1995): Weighted Averaging regression /calibration (WA) (ter Braak and van Dam 1989; Birks et al. 1990), Weighted Averaging Partial Least Square regression (WA-PLS) (ter Braak and Juggins 1993), Gaussian regression and maximum likelihood calibration (ter Braak and van Dam 1989; ter Braak et al. 1993; Vasko et al. 2000), and back-propagation (BP) (Rumelhart et al. 1986) of Artificial Neural Networks (ANNs) (Racca et al. 2001). All of these methods have inherent but different abilities to model the complex relations between taxon assemblages and environmental variables and all yield successful predictive models. However, they lack, to varying degrees, transparency as to the way information is extracted from the assemblage data and implemented in the predictive model. While it is clear that the predictive ability of these models can depend on the statistical characteristics of the calibration set (distribution and range of the environmental variable, number of samples, number of taxa, etc.), the modelling approach is also important to the final success of the model. Although some methods have been shown to outperform others in certain conditions (ter Braak and Juggins 1993; ter Braak et al. 1993; ter Braak 1995; Racca et al. 2001), little is known about the inclusion or exclusion of taxa based on their contribution to the calibration model. Generally, calibration data-sets are large and sparse and the criterion for taxon inclusion is typically *ad hoc* (e.g., all taxa with 1% relative abundance in at least one sample, present

in 2 or more samples, 1% abundance in at least two samples, 1% abundance in at least three samples, etc. (Cameron et al. 1999). Birks (1994) and Wilson et al. (1996) discuss the problem of the taxonomic precision and the optimal size of the modern data-set for the best possible calibration. Based on a series of numerical experiments with WA and $WA_{(tol)}$ models, they showed that as more taxa with low effective numbers of occurrence (N2) (Hill 1973; ter Braak 1990) are included in the modern data-set, the predictive capacity of the model increases, suggesting that (i) rare taxa contribute some ecological signal to the calibration and (ii) Hill's N2 is directly related to the contribution (in terms of predictive performance) of the taxa in these WA-based models.

Here, we suggest that the ''predictive importance'' of a given taxon in a calibration depends on what other taxa are used in the model, and hence on the extent of species redundancy in the data-set. For example, if two taxa with high N2 values have similar responses to the environmental variable inferred, the exclusion of one of them would have less impact on the calibration than the exclusion of a taxon with a different response, even if it had a low N2 value. In fact, Hill's N2 considers each taxon more or less in isolation (although relative abundance data are by definition inter-dependant) and as such it can only be viewed as an incomplete measure of importance.

In this study, we explore the question of the relative contribution of diatom taxa in the performance of palaeolimnological model calibrations. In contrast with the criterion of Hill's N2 used by Birks (1994) and Wilson et al. (1996), we suggest an alternative criterion based strictly on predictive importance. Using the Back-Propagation modelling approach of ANNs, we apply a method that (i) estimates the relevance of the diatom taxa in the calibration set and (ii) successively reduces the size of the set by excluding the least relevant taxa. For this purpose, one form of an ANN pruning algorithm, Skeletonization (Moser and Smolensky 1989), was used on the Surface Waters Acidification Programme (SWAP) diatom-pH calibration set (Birks et al. 1990). The two major objectives of this study are (i) to reduce the number of taxa contained in the modern data to test whether we can improve the predictive robustness of the model by reducing overfitting and (ii) to test whether the relative contribution of taxa in an ANN transfer-function depends on the most commonly accepted measures of importance in the palaeolimnological literature.

Methods: artificial neural networks

The neural network used here is a multi-layers perceptron trained with a back-propagation algorithm (Rumelhart et al. 1986). In this type of network, neurons are arranged in a distinct layered topology: one input layer (representing independent variables), one hidden layer, and one output layer (representing dependent variables). All neurons from one layer are connected to all neurons in the adjacent layers and all these connections have a weight that represents the parameters of the network. By back-propagation (learning process), the weights of the connections are adjusted by feeding a set of input/output pattern pairs many times. As a result of these weight adjustments, internal ''hidden'' neurons, which are not part of the input or output, come to represent important features of the task domain and the relation between input and output is captured by the interactions of these units. This relation (function) can then be used to predict output from the input data. Background information on neural networks is available in various introductory textbooks such as Bishop (1995) and more details of this methodology as applied to palaeolimnology can be found in Racca et al. (2001).

Experimental design: data-set reduction

In order to address the question of the optimal size of a calibration set and therefore the inclusion or exclusion of diatom taxa in a quantitative inference model, it is necessary to have some idea of the relative contribution of each taxon in the model's calibration.

One method for measuring the relevance ρ_i for each taxon $_i$ in the calibration set is: $\rho_i = E_{\text{without taxon}i} - E_{\text{with taxon}i}$ where E is the root mean square error (RMSE) of the model in the calibration set.

The problem with this method is that the determination of which of the N taxa should be included in the calibration set would involve the examination of 2^N possible sets of taxa. Because diatom calibration sets in palaeolimnology contain many taxa, the measure of a taxon's relevance based on the performance of every possible set is not feasible computationally. Consequently, instead of an exhaustive search, we use one form of pruning algorithm of ANNs based on an approximation of the changes in the model error function when a given taxon is removed. Pruning algorithms (see e.g., Reed (1993)) of ANNs (also called destructive algorithms) are comparable to back-

ward elimination in regression models (see e.g., Draper and Smith (1981)). Backward elimination starts with all independent variables and sequentially removes the least relevant one and stops if the model performance drops below a given threshold by the removal of any of the remaining independent variables.

The skeletonization algorithm

The skeletonization algorithm (Moser and Smolensky 1989) is used here to estimate the functionality of individual taxa in a palaeolimnological calibration model and to remove successively the least important taxa. It is a sensitivity algorithm that performs training and pruning of ANNs alternately, according to the following steps:

1. Train iteratively the network using a back-propagation function to a minimum (for details of the back-propagation algorithm, see Rumelhart et al. (1986), Racca et al. (2001));
2. Compute an approximation of the relevance ρ (for the performance of the network) of each taxon. The approximation is estimated as a first partial derivative of the error function and this derivative is computed using an error propagation very similar to that used in adjusting the weights with BP;
3. Prune the taxon with the smallest estimated relevance ρ_i;
4. Re-train the network to a minimum again (note that after deleting a taxon, the modern values of the remaining taxa are not re-expressed, so the input data are always the original relative abundance values that are used for re-training the network);
5. If the network performance (RMSE) is not higher than a certain criterion, repeat the procedure from step 2.

Details of the skeletonization algorithm are described in Moser and Smolensky (1989) and implemented in SNNS 4.2 (Stuttgart Neural Network Simulator, Zell et al. (1996)).

Validation of the reduced models

Because skeletonization pruning of taxa is based on the change of the error function in the calibration set (apparent RMSE), after the size reduction of the data-set, a validation is made using a standard back-propagation model with cross-validation based on leave-one-out jackknifing. For this purpose, the same methodology as proposed in Racca et al. (2001) was applied using a cross-validation routine (CROSVAL, Racca, unpublished program) of YANNS (Yet Another Neural Network Simulator, Boné et al. (1998)).

Data-set

The SWAP diatom-pH data-set used here is the same as used by Birks et al. (1990) and described by Stevenson et al. (1991) and includes all taxa that are present in at least two samples with an abundance of 1% or more in at least one sample. The diatom data-set is summarized in terms of number of samples, number of taxa, percentage number of non-zero values (occurrence), the total inertia (variance), the range, mean, and median of the effective number (N2) of taxa per sample and the effective number of occurrences (N2) of each taxon. The modern pH values are summarized in terms of the range, mean, median, and standard deviation (Table 1).

Results and discussion

Skeletonization pruning and ANN model performance

The SWAP diatom data-set was pruned according to

Table 1. Descriptive statistics for the SWAP diatom-pH data set

		Minimum	Median	Mean	Maximum	S.D	Range
Number of samples	167						
Number of taxa	267						
% number of positive values in data	18.47						
Total inertia	3.39						
N2 for samples		5.13	28.58	29.22	57.18		
N2 for taxa		1	14.99	23.76	120.86		
pH		4.33	5.27	5.56	7.25	0.77	2.92

the relevance of the diatom taxa to the performance of the model. Although the algorithm prunes the species data-set taxon by taxon, taxa were grouped in classes of importance (each containing 5% of the total taxa). This is because (i) the order of relevance, and therefore of deletion, can be slightly different depending on the initial parameters of the network and the pruning process (weight initialization, learning rate, number of iterations on training and re-training steps) and because (ii) no validation is made during the pruning steps (the validation of *N* new neural networks models is extremely time consuming).

The skeletonization algorithm allowed us to reduce the data-set by 85% (267 to 37 taxa), without significantly affecting the model's performance ($RMSEP_{jackknife}$ of 0.323 to 0.334, $r^2_{jackknife}$ 0.82 to 0.81, $F = 0.93$ $p > 0.05$). This is a remarkable result given that the literature suggests that all taxa are important and should be retained (Birks 1994). The pruning gave 17 classes of taxa of increasing importance (Table 2). Interestingly, while the cross-validated model performance remained unchanged (jackknifed r^2 or RMSEP), the species reduction resulted in a significant decrease in the *apparent* performance of the calibration set (apparent RMSE of 0.163 to 0.285, apparent r^2 of 0.96 to 0.86, $F = 2.98$ $p < 0.005$) (Table 2). Figure 1 illustrates how the difference between the apparent RMSE and $RMSEP_{jackknife}$ decreases with the number of taxa in the training set. Ideally, the apparent RMSE of any model should be a reliable measure of the actual predictive power of a model and any difference between the apparent and cross-validated measures indicates the extent to which the model overfitted the data. Overfitted models also have a tendency to reduce the robustness of the model, i.e., to be more sensitive to small changes in model input values. This implies that even if the removal of many taxa (taxa with redundant information, for example) does not greatly improve or deteriorate the predictive performance of the model as

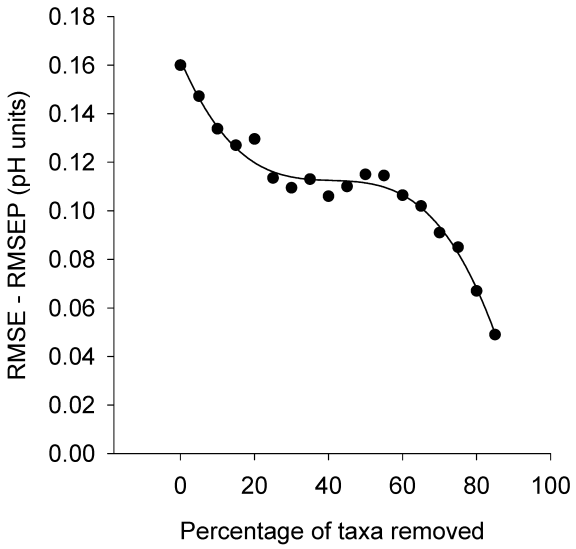

Figure 1. Changes in the difference between root mean square error of estimation (apparent RMSE) and root mean square error of prediction ($RMSEP_{jackknife}$) with an increasing number of taxa removed according to their functionality.

Table 2. Summary statistics of the SWAP diatom pH inference models according to the classes of taxa included based on the Skeletonization procedure

Class of importance	Size reduction (%)	RMSE	RMSEP	Mean bias	Max bias
	0	0.163	0.323	−0.030	−0.441
1	5	0.174	0.322	−0.040	−0.467
2	10	0.182	0.323	−0.037	−0.475
3	15	0.198	0.325	−0.039	−0.506
4	20	0.199	0.327	−0.044	−0.498
5	25	0.214	0.328	−0.047	−0.512
6	30	0.215	0.325	−0.043	−0.509
7	35	0.213	0.326	−0.042	−0.511
8	40	0.218	0.324	−0.037	−0.513
9	45	0.219	0.329	−0.035	−0.501
10	50	0.208	0.323	−0.033	−0.499
11	55	0.214	0.329	−0.031	−0.467
12	60	0.216	0.328	−0.024	−0.455
13	65	0.221	0.330	−0.037	−0.540
14	70	0.240	0.331	−0.035	−0.540
15	75	0.256	0.336	−0.032	−0.562
16	80	0.269	0.335	−0.021	−0.418
17	85	0.285	0.334	−0.013	−0.460

assessed by $RMSEP_{jackknife}$, it does help minimize overfitting. However, it is surprising and difficult to explain why this decrease is not linear.

Further analysis of the performance of each pruned model was done by comparing the average bias and the maximum bias in the prediction model (ter Braak and Juggins 1993). For estimation of the 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 and Juggins 1993). Table 2 shows the statistics for each reduced model. As no difference is observed for the mean or maximum bias between the initial set (all taxa) and all the reduced sets, it appears that the deletion of taxa based on Skeletonization pruning does not affect the general trends of these models.

Even if the results show that the predictive capacity of all reduced models is similar in global terms such as $RMSEP_{jackknife}$, $r^2_{jackknife}$ and bias, the reconstructions for a given lake may be different between the reduced models (Figure 2). Thus, the question remains as to what extent can these various differences in prediction be attributed to differences in taxon inclusion or whether a model built using only selected taxa can produce more reliable predictions when applied to fossil down-core reconstructions. We examined this question by comparing down-core pH reconstructions for the Round Loch of Glenhead using various levels of skeletonization. Each ANN-based skeleton inference model was applied to the 101 fossil assemblages from the Round Loch of Glenhead core, representing 10 300 radiocarbon years of continuous sediment deposition (Jones et al. 1989). ANN skeleton-based-pH-reconstruction with 30%, 60%, and 85% pruned taxa (Figure 3a-c) were compared to the reconstruction based on the full-set ANN, WA, and ML based models (Figure 3d-f). All these transfer functions produce a generally similar pH reconstruction for the lake (Figure 3a-f), suggesting that (i) the inclusion of more than 15% of taxa does not provide more information about past pH levels and (ii) even when using this 15% of taxa for calibration, ecologically important taxa present in the core sample are also contained in the calibration. However, it is surprising and somewhat disturbing that it is when *all* taxa are included in the calibration that the pH reconstructions based on ANN and WA predictions differ the most and, for some periods, this difference even exceeds the associated error (RMSEP) of each model. We suggest that this may due to the greater sensitivity of overfitted models and that more reliable predictions may actually be obtained from the pruned model. Until independent evidence is found, however, this will remain a conjecture. Nonetheless, our results show that, for reconstruction purposes, a model cannot be judged solely on global measures of predictive power like RMSEP (since both full-data ANN and WA models had equivalent RMSEP). On palaeolimnological grounds, the reconstructed pH values below 5 between 2500 and 8500 B.P. suggested by the full-set ANN model (Figure 3d) seem unlikely in the absence of the input of strong acids prior to the Industrial Revolution. We suggest that greater efforts should be made towards the development of reliable and independent measures of robustness. Until such tools are developed, however, our results argue that non-essential taxa may only improve the *apparent* predictive power, and may in fact be a detriment to down-core reconstructions.

Comparison between Skeletonization pruning and N2 selection on model performance

To compare the Skeletonization pruning procedure with the deletions used by Birks (1994) and Wilson et al. (1996), new ANN models were constructed using the SWAP data-set successively reduced by 5% groups in decreasing order of Hill's N2 value. Their

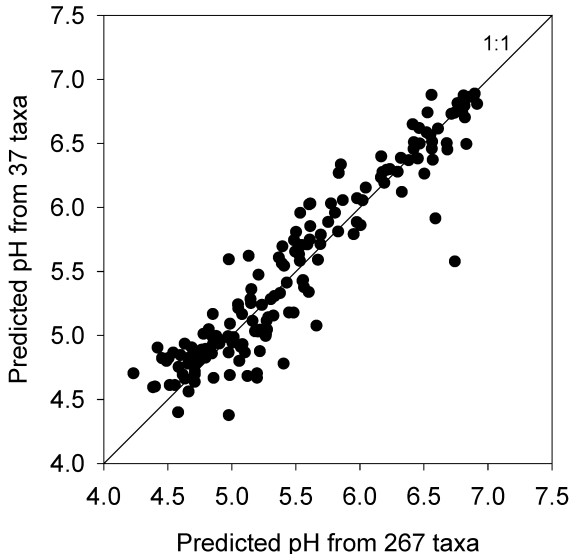

Figure 2. Plot illustrating the differences between predictions (leave-one out jackknifing predictions) for each lake in the calibration data-set depending on the number of taxa included in the model. The fitted line is a 1:1 line.

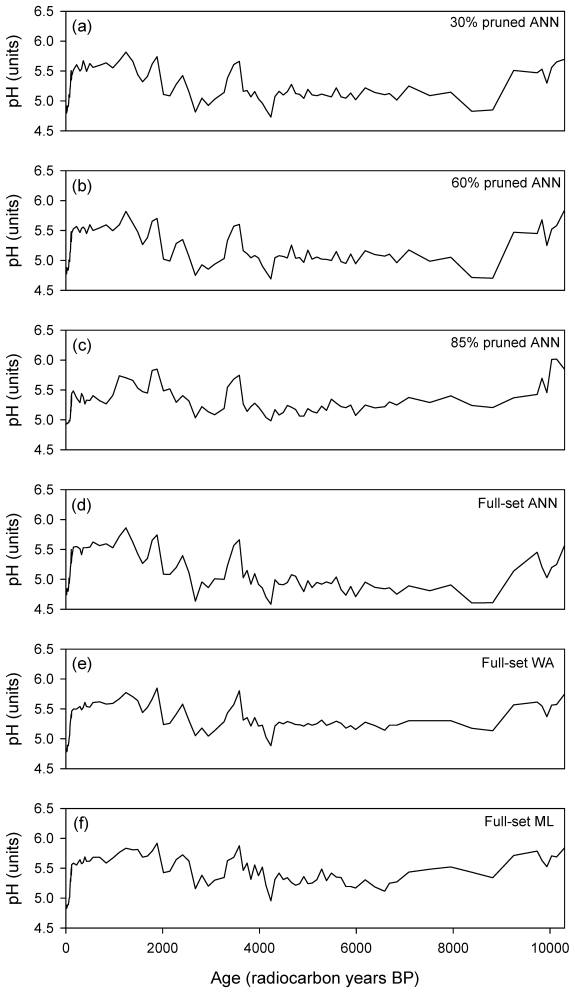

Figure 3. Diatom-inferred pH at the Round Loch of Glenhead (UK) for the past 10,300 radiocarbon years. based on (a) 30% pruned ANN based model, (b) 60% pruned ANN based model, (c) 85% pruned ANN based model, (d) full-set ANN based model, (e) full-set WA based model, and (f) full-set Gaussian maximum likelihood based model. The new WACALIB program (version 3.5) with the de-bug in the maximum likelihood algorithm (Birks 2001) was used here.

respective predictive capabilities were then compared using leave-one-out jackknifing on all the reduced taxa data-sets. Figure 4 shows the changes of $RMSEP_{jackknife}$ when the data-sets are reduced by the two methods. When taxa are removed according to their N2 values, only 20% of taxa can be excluded, as opposed to 85% when using skeletonization, as model performance starts to show a major decrease once 25% of taxa were excluded ($RMSE_{jackknife}$ of 0.323 to 0.343). This decrease progressively continues to reach an $RMSEP_{jackknife}$ of 0.430 ($r^2_{jackknife}$ of 0.69) when 85% of taxa are removed. On the basis of the N2 values, our results corroborate the findings of Birks (1994) who showed that, for the same data-set, the lowest $RMSE_{boot}$ of prediction (0.32) occurs in WA regression and calibration prior to deletion of any taxa. However, even if the performance of WA models depends on the N2 value of the taxa, the conclusion from this series of numerical experiments is that the functionality of individual taxa in an ANN model calibration is not related to their N2 values. Also, in a WA model, when only using the 37 selected taxa on the basis of their ANN functionality, the performance ($RMSEP_{jackknife}$ of 0.424, $r^2_{jackknife}$ of 0.70) is similar to that obtained with the 37 highest N2 taxa ($RMSEP_{jackknife}$ of 0.411, $r^2_{jackknife}$ of 0.72). These results imply that the functionality (and possibly the importance) of a given taxon can be different between an ANN and a WA model. This supports our previous suggestion (Racca et al. 2001), that the two modelling approaches are both conceptually and mathematically different: a taxon with a high functionality in an ANN model may have a low functionality in a WA model and vice versa.

Comparison between skeletonization-based and other measures of taxon importance

As the relative contribution estimated by the ANN pruning algorithm can be biased for correlated taxa,

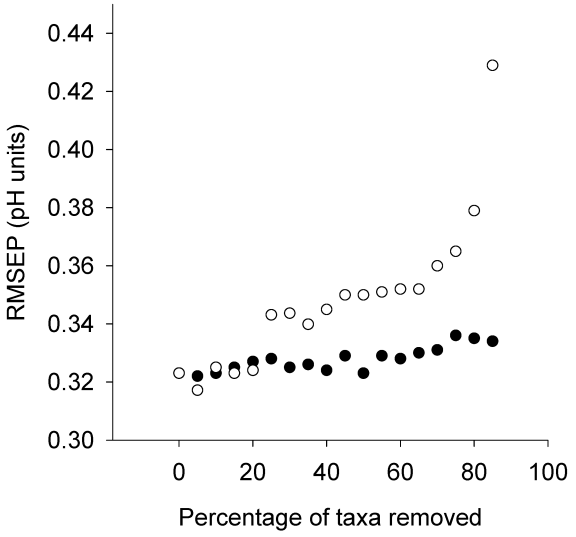

Figure 4. Plot illustrating the changes of the root mean square error of prediction ($RMSEP_{jackknife}$) when taxa are removed according to their N2 values (white circles) and to their functionality (black circles).

we suggest exercising prudence when comparing the effective (i.e., skeletonization-based) importance of taxa with other commonly believed (marginal) measures of importance. For this reason, we made no attempt to compare the individual taxon's level of importance. However, it is possible to compare the marginal characteristics of each pruned group (5%) of taxa. Figure 5 shows the distribution of these characteristics. These characteristics are grouped on the box plots in term of maximum, minimum, median, and inter-quartile range of Hill's N2 (Figure 5c), number of occurrences (Figure 5b), and WA tolerance (Figure 5a). If marginal importance was truly correlated to effective importance in an ANN model calibration, then the pruning process used here should have removed the least marginally important taxa first. However, this is not the case, suggesting that effective importance is not driven by any of the commonly regarded measures of importance (N2, number of occurrences, WA tolerance). As demonstrated above, this also seems to be true for WA models: their performances are similar using the 37 most functional taxa (for ANN; Table 3) or the 37 taxa with the highest N2. Therefore, we suggest that for both an ANN or WA approach, these measures of marginal importance may be of little practical use other than for a preliminary description of the data: a taxon with a high marginal importance may have a low causal and predictive importance, and a taxon with a low marginal importance may have a high causal and predictive importance. However, it should be stressed that the taxa selected on the basis of their effective importance in an ANN model cannot necessarily be used to construct a reliable WA model.

Table 3. General characteristics of the 37 most functional taxa for calibration based on ANN modelling approach

Species	Number of occurrences	WA tolerance	Hill's N2
Achnanthes minutissima minutissima (Kutz., 1833)	118	0.672	46.062
Achnanthes marginulata (in Cleve & Grun., 1880)	118	0.618	30.867
Asterionella formosa formosa (Hassall, 1850)	27	0.363	11.309
Asterionella formosa ralfsii (W. Sm.) Wolle, 1890	11	0.694	3.448
Actinocyclus normanii normanii (Greg. ex Grev.)	13	0.397	8.029
Aulacoseira subborealis (SWAP 1989)	4	0.451	1.373
Aulacoseira sp form 1	29	0.705	16.828
Aulacoseira sp form 2	13	0.323	6.417
Brachysira vitrea (Grun.)	140	1.159	50.176
Cymbella cesatii cesatii (Grun.)	23	0.759	7.644
Cymbella gracilis (Rabenh.) Cleve, 1894	118	0.733	60.512
Cymbella gracilis scotica (W. Sm.) Rabenh., 1864	7	0.840	3.925
Cyclotella comta comta (Ehrenb.) Kutz., 1849	39	0.388	20.442
Cyclotella comensis (Grun.) in Van Heurck, 1882	29	0.590	6.363
Cyclotella sp	101	0.966	24.939
Eunotia exigua tridentula	10	0.334	6.016
Eunotia diodon (Ehrenb., 1837)	18	0.560	11.070
Eunotia parallela parallela (Ehrenb., 1843)	9	0.372	4.020
Eunotia incisa (W. Sm.) ex Greg., 1854	145	0.554	70.353
Eunotia curvata subarcuata (Naegeli ex Kutz.)	33	0.479	20.510
Fragilaria construens venter (Ehrenb.)	49	0.893	23.225
Fragilaria elliptica (Schum., 1867)	30	0.716	7.139
Gomphonema angustatum angustatum (Kutz.)	31	0.770	13.104
Navicula pupula pupula (Kutz., 1844)	47	0.850	17.242
Navicula angusta (Grun., 1860)	56	0.949	17.773
Navicula soehrensis soehrensis (Krasske, 1923)	13	0.234	1.559
Navicula impexa (Hust., 1961)	21	0.951	6.066
Navicula bacilliformis (Grun.) in Cleve & Grun., 1880	6	0.312	1.385
Navicula seminuloides (Hust., 1937)	9	0.630	3.679
Nitzschia perminuta (Grun.) in Van Heurck	19	0.315	10.843
Nitzschia palea tenuirostris (Grun.) in Van Heurck, 1881	9	0.484	4.921
Nitzschia acula (Hantzsch) ex Cleve & Grun., 1880	14	0.709	8.046
Pinnularia microstauron microstauron (Ehrenb.) Cleve, 1891	80	0.733	35.718
Surirella linearis linearis (W. Sm., 1853)	39	0.889	11.143
Tabellaria fenestrata (Lyngb.) Kutz., 1844	18	0.932	10.626
Tabellaria quadriseptata (Knudson, 1952)	87	0.614	23.085

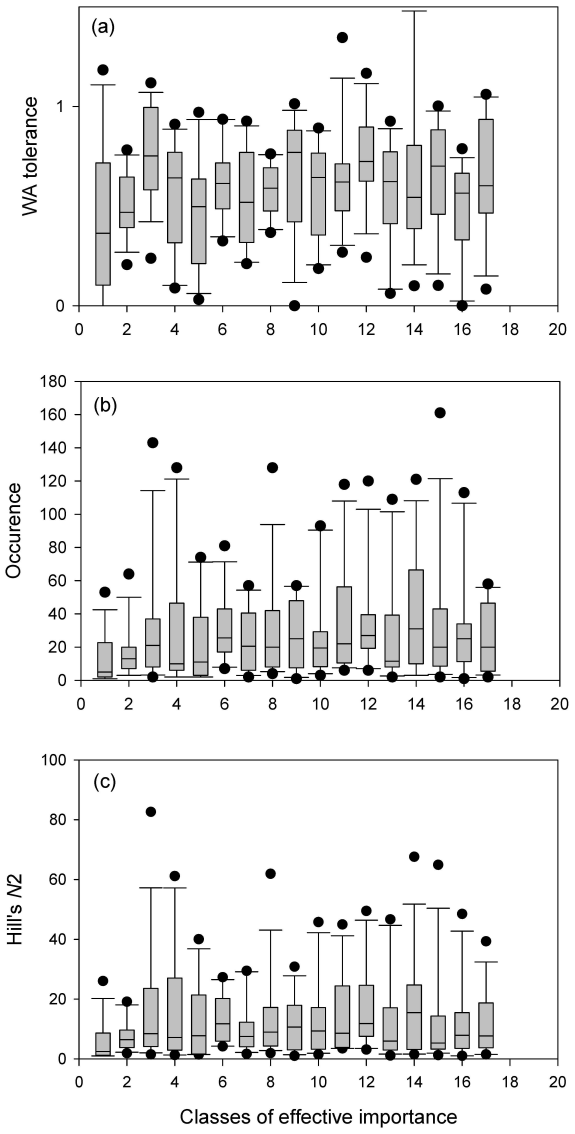

Figure 5. Box plots of the general characteristics of (a) WA-tolerance, (b) number of occurrences and (c) Hill N2 values for each class of taxon importance.

General discussion

Taxa contribution and choice of method for taxon inclusion

In theory, the inclusion of all taxa is desirable in a down-core reconstruction because it increases the probability that the fossil taxa will also be found in the modern calibration set. On the other hand, to justify the inclusion of all taxa in the construction of a transfer function model, one must assume that they all contribute, albeit to different extents, to the true (not just the apparent) performance of the model. However, our study suggests that this is not the case. Instead, it is clear that some taxa have less importance in the calibration and thus can be ignored in the model. These taxa will surely differ, depending on the variable being studied, as their effective relevance will vary depending on the environmental variable of interest. Therefore, any measure which is totally independent of the environmental variable, such as N2 or number of occurrences, cannot be considered a complete measure of relevant importance and should therefore not be used as a criterion for eliminating taxa. Intuitively, this suggests that a measure which does depend on the environmental variable, such as WA tolerance, should be a more useful way of evaluating effective importance. Again, our data suggest it is not the case (Figure 5a).

Although tolerance can easily be fitted into a coherent ''ecological'' theory of importance, it is apparently of little empirical value to the predictive power of the calibration model. Therefore, while it may be appropriate to use an environmentally dependent measure (such as tolerance) to eliminate taxa from a data-set, this criterion does not guarantee its reliability either. We suggest that the exclusion of taxa based on a strictly empirical measure of importance is attractive for at least two reasons. First, the functionality estimation of an individual taxon based on a pruning procedure is environmentally dependent. Second, in this type of approach, the selection of taxa is made in order to create an optimal model for a given task with no *a priori* knowledge (based only on a fixed level of error) whereas in other approaches, the selection of taxa is based on *a priori* knowledge in order to create an optimal solution. Therefore, even if the inclusion of all taxa produces reasonable results, we suggest that it may only be a sub-optimal solution.

Optimal models and taxa data-set reduction

Because calibration sets in palaeolimnology often contain many taxa and few lakes, all modelling techniques proposed for developing quantitative inference models suffer to a large degree from the ''curse of dimensionality'' (ter Braak 1995). This translates into transfer functions that typically overfit the relationship between taxon assemblages and environmental variables. This is why ''brute-force'' procedures such

as jackknifing and bootstrapping were introduced in palaeolimnology in the first place: to minimise the problem of overfitting and to obtain more realistic measures of predictive power (Birks 1995, 1998). However, it should be borne in mind that, ultimately, it is always an overfitted model that is used in reconstruction. Only its purported predictive power has been estimated and toned down by jackknifing or bootstrapping. We contend that this is far from the best modelling approach. Instead, we suggest that a more optimal model should contain all but also only those taxa that are necessary for the model to perform well, and we therefore need efficient methods to separate the ''wheat from the chaff''. Only when the clutter from our overburdened palaeolimnological inference models has been removed can we hope to improve our understanding of the model by finding, for example, the ecological characteristics of the taxa that are deemed important.

Although it would seem obvious that the ''curse of dimensionality'' is directly related to the ratio of the number of taxa to the number of lakes (as this ratio determines the ratio of the dimensional space in which the function is determined to the number of observations for which the function is determined), no studies seem to have taken this into account. However, when we compiled the predictive characteristics of 35 recent palaeolimnological diatom-based inference models (Table 4), it is clear that the degree of generalization (robustness) of these models is inversely related to their dimensionality (Figure 6). This robustness is expressed here as the ratio between the apparent performance (RMSE) and the cross-validated performance (RMSEP) (Figure 6). Only when

Table 4. Examples of the recently published diatom-based inference models in paleolimnology used in Figure 6

References	Env. variables	Samples	Taxa	Models
Joynt and Wolfe (2001)	Air temperature	61	107	WA
Korsman and Birks (1996)	Alkalinity	119	115	WA
Joynt and Wolfe (2001)	Conductivity	61	107	WA
Ng and King (1999)	Conductivity	93	53	WA
Reed (1998)	Conductivity (log 10)	74	169	WA
Korsman and Birks (1996)	Colour (log 10)	119	115	WA
Moser et al. (2000)	Depth	35	112	WA
Moser et al. (2000)	Depth	53	177	WA-PLS (3)**
Dixit et al. (1993)	DOC	71	188	WA
Philibert and Prairie (2002)	DOC	41	160	WA-PLS (2)**
Philibert and Prairie (2002)	DOC	35	101	WA
Philibert and Prairie (2002)	DOC	76	214	WA-PLS (2)**
Rosen et al. (2000)	pH	50	157	WA
Cameron et al. (1999)	pH	118	530	WA-PLS (3)**
Cameron et al. (1999)	pH	167	277	WA-PLS (2)**
Racca et al. (2001)	pH	76	214	WA-PLS (3)**
Hall and Smol (1996)	pH	54	92	WA
Joynt and Wolfe (2001)	pH	61	107	WA
Dixit et al. (1993)	pH	71	188	WA
Philibert and Prairie (2002)	pH	41	160	WA-PLS (3)**
Philibert and Prairie (2002)	pH	35	101	WA-PLS (2)**
Korsman and Birks (1996)	pH	119	115	WA
Wilson et al. (1994)	Salinity	102	107	WA
Wilson et al. (1994)	Salinity	42	54	WA
Roberts and Mcminn (1998)	Salinity (log 10)	33	47	WA
Roberts and Mcminn (1998)	Salinity (log 10)	33	47	WA-tol
Rosén et al. (2000)	Surface temp (J)*	52	157	WA-PLS (3)**
Lotter et al. (1997)	Surface temp (S)*	64	345	WA-PLS (2)**
Pienitz et al. (1995)	Surface temp (S)*	61	107	WA
Rosén et al. (2000)	TOC	33	157	WA-PLS (3)**
Hall and Smol (1996)	TP	54	92	WA
Reavie et al. (1995)	TP	64	150	WA
Lotter et al. (1998)	TP (log 10)	72	341	WA-PLS (2)**

*(S)=Summer, (J)=July **number of components, WA – Weighted averaging, WA-tol – Tolerance-Downweighted weighted averaging, WA-PLS – Weighted averaging partial least squares, DOC – Dissolved organic carbon, Temp – Temperature, TOC – Total organic carbon, TP – Total phosphorus

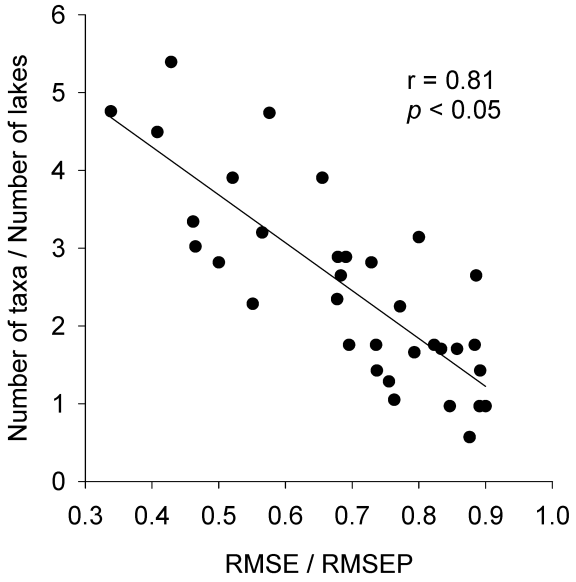

Figure 6. Relationship between the robustness (RMSE/RMSEP) of a model and the ratio of the number of taxa to the number of lakes in 35 published palaeolimnological diatom-based inference models.

the apparent and cross-validated predictive powers are similar can we safely say that the model is robust. The implications of this relationship are obvious; in order to achieve maximum robustness, the ratio of taxa : lakes must be decreased to a minimum. This can be accomplished by one of two ways, increase the number of lakes or lower the number of taxa. However, increasing the number of lakes would inevitably lead to an increase in the number of taxa, thereby defeating the purpose. Moreover, increasing the number of lakes is often not possible due to logistical constraints. The remaining solution is to diminish the number of taxa. The skeletonization procedure proposed here seems to accomplish this efficiently, resulting in an ANN model that offers both robustness and predictive power. Such a model may help attain the two main goals of quantitative palaeolimnology namely (i) to develop robust calibration models with the highest predictive power possible and (ii) to develop robust transfer functions capable of reliably reconstructing palaeoenvironments.

Conclusions

The results clearly show that, for calibration purposes, most taxa are not needed and that their usefulness in a calibration model is not correlated to measures such as N2, number of occurrences, or tolerance. Therefore, in order to build efficient ANN palaeolimnological transfer-function models, taxa with little or no effective importance need to be removed. The procedure proposed here is one way of achieving this, resulting in a more robust model, with at least as much predictive power as other more complex models. Once validated using other data-sets, this procedure could prove a very useful tool for palaeolimnological reconstruction based on ANNs. It is also hoped that the approach can help shed light on what makes a particular taxon important and another taxon not important in a palaeolimnological reconstruction.

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 are grateful to R. W. Battarbee for giving permission to use the SWAP modern diatom-pH data and to V. J. Jones for providing her Round Loch of Glenhead diatom core-data. We also thank A.F. Lotter for his constructive comments on the manuscript.

References

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

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

Birks H.J.B. 1998. Numerical tools in palaeolimnology-Progress, potentialities, and problems. J. Paleolim. 20: 307–332.

Birks H.J.B. 2001. Maximum likelihood environmental calibration and the computer program WACALIB- a correction. J. Paleolim. 25: 111–115.

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

Bishop C.M. 1995. Neural Networks for Pattern Recognition. Oxford Clarendon Press, Oxford, pp. 482.

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

Cameron N.G., Birks H.J.B., Jones V.J., Berge F., Catalan J., Flower R.J et al. 1999. Surface-sediment and epilithic diatom pH calibration sets for remote European mountain lakes (Al:PE Project) and their comparison with the surface waters acidifica-

tion programme (SWAP) calibration set. J. Paleolim. 22: 291–317.

Dixit S.S., Cumming B.F., Birks H.J.B., Smol J.P., Kingston J.C., Uutala A.J. et al. 1993. Diatom assemblages from Adirondack lakes (New York, USA) and the development of inference models for retrospective environmental assessment. J. Paleolim. 8: 27–47.

Draper N.R. and Smith H. 1981. Applied Regression Analysis. 2nd edn. Wiley, New York.

Hall R.I. and Smol J.P. 1996. Paleolimnological assessment of long-term water-quality changes in south-central Ontario lakes affected by cottage development and acidification. Can. J. Fish. Aquat. Sci. 53: 1–17.

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

Jones V.J., Stevenson A.C. and Battarbee R.W. 1989. Acidification of lakes in Galloway, south west Scotland: a diatom and pollen study of the post-glacial history of the Round Loch of Ghenhead. J. Ecol. 77: 1–23.

Joynt E.H. and Wolfe A.P. 2001. Paleoenvironmental inference models from sediment diatom assemblages in Baffin island lakes (Nunavut, Canada) and reconstruction of summer water temperature. Can. J. Fish. Aquat. Sci. 58: 1222–1243.

Korsman T. and Birks H.J.B. 1996. Diatom-based water chemistry reconstructions from northern Sweden: A comparison of reconstruction techniques. J Paleolim 15: 65–77.

Lotter A.F., Birks H.J.B., Hofmann W. and Marchetto A. 1997. Modern diatom, cladocera, chironomid, and chrysophyte cyst assemblages as quantitative indicators for the reconstruction of past environmental conditions in the Alps. I. Climate. J. Paleolim. 18: 395–420.

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

Moser K.A., Korhola A., Weckstrom J., Blom T., Pienitz R., Smol J.P. et al. 2000. Paleohydrology inferred from diatoms in northern latitude regions. J. Paleolim. 24: 93–107.

Moser M. and Smolensky P. 1989. Skeletonization: A technique for trimming the fat from a network via relevance assessment. Advances in Neural Information Processing Systems (NIPS) 1: 107–115.

Ng S.L. and King R.H. 1999. Development of a diatom-based specific conductivity model for the glacio-isostatic lakes of Truelove Lowland: Implications for paleoconductivity and paleoenvironmental reconstructions in Devon Island lakes, NWT, Canada. J. Paleolim. 22: 367–382.

Philibert A. and Prairie Y.T. 2002. Diatom-based transfer functions for western Quebec lakes (Abitibi and Haute Maurice): the possible role of epilimnetic CO_2 concentration in influencing diatom assemblages. J. Paleolim. 27: 465–480.

Pienitz R., Smol J.P. and Birks H.J.B. 1995. Assessment of freshwater diatoms as quantitative indicators of past climatic-change in the Yukon and Northwest Territories, Canada. J. Paleolim. 13: 21–49.

Racca J.M.J., Philibert A., Racca R. and Prairie Y.T. 2001. A comparison between diatom-pH-inference models using Artificial Neural Networks (ANNs), Weighted Averaging (WA) and Weighted Averaging Partial Least Square (WA-PLS) regressions. J. Paleolim. 26: 411–422.

Reavie E.D., Hall R.I. and Smol J.P. 1995. An expanded weighted-averaging model for inferring past total phosphorus concentrations from diatom assemblages in eutrophic British-Columbia (Canada) lakes. J. Paleolim. 14: 49–67.

Reed J.M. 1998. A diatom-conductivity transfer function for Spanish salt lakes. J. Paleolim. 19: 399–416.

Reed R. 1993. Pruning algorithms - a Survey. IEEE Transactions on Neural Networks 4: 740–747.

Roberts D. and Mcminn A. 1998. A weighted-averaging regression and calibration model for inferring lake water salinity from fossil diatom assemblages in saline lakes of the Vestfold Hills: a new tool for interpreting Holocene lake histories in Antarctica. J. Paleolim. 19: 99–113.

Rosén P., Hall R., Korsman T. and Renberg I. 2000. Diatom transfer-functions for quantifying past air temperature, pH and total organic carbon concentration from lakes in Northern Sweden. J. Paleolim. 24: 109–23.

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

Stevenson A.C., Juggins S., Birks H.J.B., Andreson D.S., Andreson N.J., Battarbee R.W. et al. 1991. The Surface Waters Acidification Project Palaeolimnolgy Programme: Modern Diatom/Lake-Water Chemistry data-set. Ensis Publishing, London, 86 pp.

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

ter Braak C.J.F. 1995. Nonlinear methods for multivariate statistical calibration and their use in paleoecology – A comparison of inverse (k-nearest neighbors, partial-least squares and weighted averaging partial least squares) and classical approaches. Chemometrics and Intelligent Laboratory Systems 28: 165–180.

ter Braak C.J.F. and Juggins S. 1993. Weighted Averaging Partial Least-Squares Regression (WA-PLS) - an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia 269: 485–502.

ter Braak 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. In: Patil G.P. and Rao C.R. (eds), Multivariate Environmental Statistics. Elsevier Science Publishers, Amsterdam, pp. 525–560.

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

Vasko K., Toivonen H.T.T. and Korhola A. 2000. A Bayesian Multinomial Gaussian Response Model for organism-based environmental reconstruction. J. Paleolim. 24: 243–250.

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

Wilson S.E., Cumming B.F. and Smol J.P. 1994. Diatom-salinity relationships in 111 lakes from the Interior Plateau of British Columbia, Canada: the development of diatom-based models for paleosalinity reconstructions. J. Paleolim. 12: 197–221.

Zell A., Mamier G., Vogt M., Mache N., Hübner R., Döring S. et al., Stuttgart Institute for Parallel and Distributed High Performance Systems University of Stuttgart 1996. Stuttgart Neural Network Simulator v4. 2. ftp/informatik/uni-stuttgart/de/in pubSNNS SNNSv4. 1.