J Paleolimnol (2007) 38:467–472
DOI 10.1007/s10933-006-9082-x

NOTE

PaleoNet: new software for building, evaluating and applying neural network based transfer functions in paleoecology

Julien M. J. Racca · Robert Racca · Reinhard Pienitz · Yves T. Prairie

Received: 24 June 2006 / Accepted: 1 December 2006 / Published online: 30 January 2007

Abstract Transfer functions that implement organism–environment relationships are now commonly used for inferring past environmental conditions in paleoecology. Specific software for developing and evaluating commonly used modelling techniques such as Weighted averaging (WA), Weighted averaging partial least square (WA-PLS), Maximum likelihood (ML), and Modern analog technique (MAT) are available. A new software programme, *PaleoNet*, is now available for modelling organism–environment relationships which is specifically designed for the development and the evaluation of artificial neural network (ANN) based transfer functions in paleoecology. Here we present the main characteristics of this new software *PaleoNet* (User guide version 1.01) and discuss in more detail one of its specific features: the pruning.

Keywords Artificial neural networks · Paleo-reconstruction · Pruning · Software

J. M. J. Racca (✉) · R. Pienitz
Paleolimnology-Paleoecology Laboratory, Centre d'Études Nordiques & Département de Géographie, Université Laval, Québec, QC, Canada G1K 7P4
e-mail: racca.julien@courrier.uqam.ca

R. Racca
Département des sciences, Université de Nouvelle Calédonie, Site de Nouville, BP 4477, 98847 Nouméa Cedex, Nouvelle Calédonie
e-mail: racca@univ-nc.nc

Y. T. Prairie
Département des Sciences Biologiques, Université du Québec à Montréal, succ. Centre-Ville, Case postale 8888, Montréal, QC, Canada H3C 3P8
e-mail: prairie.yves@uqam.ca

Introduction

The necessity for hindcasting ecosystem dynamics and disturbances for periods where instrumental data are non-existent has stimulated a great number of efforts for the development and evaluation of analytical inference methods over the past 30 years. Beginning with the work of Imbrie and Kipp (1971), over 1,000 local or regional transfer functions for quantifying "organism–environment" relationships have been developed and used within the framework of paleoecological studies (H.J.B. Birks, personal communication). Generally based on the calibration of the organisms' response curve relative to one environmental gradient, such transfer functions are used for inferring past environmental conditions from the composition and the abundance of fossil assemblages. Various numerical approaches (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), Bayesian method (Vasko et al. 2000), and the backpropagation (Rumelhart et al. 1986) of artificial neural networks (ANN; Racca et al. 2001), were proposed. Although it is possible to determine the main theoretical advantages and disadvantages of each method, the predictive capacity of the resulting models is often similar or at least difficult to foresee. Various software intended specifically for paleoecological applications are available and widely used (e.g. CALIBRATE, Juggins and ter Braak 1993; WA-PLS, Juggins and ter Braak 1995; C2, Juggins 2004). Although these software packages implement the majority of the algorithms commonly used, none proposes the backpropagation of ANNs. However, some theoretical properties of the backpropagation algorithm make this modelling approach very useful, especially when complex non-linear or non-unimodal relations, such as those often seen in paleoecology, have to be implemented. Moreover, some very useful tools of the ANN approach, such as the "pruning" (e.g. Reed 1993; Thimm and Fiesler 1997), have not been implemented in more common numerical methods yet. Here, we present new software for building, evaluating and applying neural network based transfer functions in paleoecology. *PaleoNet* is a simulator of artificial neural networks intended for researchers within the paleoscience community. It is an autonomous software developed under Matlab. It is divided into two principal parts: the first is intended for the development and the evaluation of transfer functions, whereas the second focuses on their application. *PaleoNet* and its user's guide are available free of charge from the following address: http://www.cen.ulaval.ca/paleo/.

Building artificial neural network transfer functions with *PaleoNet*

The development of an artificial neural network transfer function in paleoecology is, in a mathematical sense, a problem of function approximation. Starting from an organism–environment calibration data set, the problem consists primarily of finding the mathematical function (model) that can be used to predict (with a minimal average error) one environmental gradient using the organism/species data (distribution and abundance). This can be achieved by various gradient descendent algorithms of artificial neural networks such as backpropagation (Rumelhart et al. 1986). Gradient descendent algorithms imply an iterative training mechanism by which the function approximation is carried out. In this case this iterative training process is said to be "supervised" because the function approximation is based on existing examples (i.e., the modern calibration or training data-set). Supervised training (called also learning process) is applied to a network architecture and is controlled by some initial training parameters such as the "learning rate" and the "number of learning cycles". To obtain optimal training, the network architecture and the training parameters need to be adjusted before running a simulation. Artificial neural networks have recently been applied to a number of paleoecological studies (e.g. Malmgren and Nordlund 1997; Malmgren et al. 2001; Kucera et al. 2005; Racca et al. 2004).

PaleoNet is a simulator of multi-layer perceptron. A multi-layer perceptron is a network whose architecture is made of successive connected layers of artificial neurons. *PaleoNet* implements networks with three layers. In this type of architecture, the first layer (network input) serves to introduce the organism/species data while the third layer (network output) represents the environmental gradient to be modeled. The intermediate layer (called the hidden layer) serves to indirectly connect the inputs to the output of the network. While the number of artificial neurons in the input and output layers is defined by the characteristics of the data-set (mainly by the number of species), the appropriate number of neurons in the hidden layer is more difficult to define a priori. In fact, although the function approximation capability depends, in part, on the number of hidden neurons, there is no formula enabling us to determine its optimal number. However, in general the larger the number of neurons in the intermediate layer, the easier the network will converge during the training process.

On the other hand, the generalization capability is less accurate when the number of neurons on the intermediate layer is too large, and vice versa. Thus, the number of neurons on the hidden layer must be determined in order to obtain a balanced adjustment between the capacities to converge and to generalize. *PaleoNet* allows the user to define the number of neurons on the hidden layer.

Ultimately, if the number of neurons in the hidden layer is large and the number of training cycles long enough, then the network will fit any data perfectly. This is why "backpropagation networks" are also called "universal approximators". Thus, these networks are only really useful if they are capable, after a learning period, of generalizing (i.e. of predicting reliably for observations not included in the data set). A well-designed neural network will, after training with a learning set, give correct predictions when fed a validation set. Several types of validation techniques exist but the most commonly used in paleoecology involve jack-knifing or *K*-fold cross-validation principles. However, Telford and Birks (2005) found that transfer functions based on ANN are likely to over-fit the data unless a spatially independent test set is available to help model selection. They found that models based on unimodal species–environment response (WA and WA-PLS) are more robust to spatial autocorrelation of environmental variables in the training set than ANN models, suggesting that jack-knifing cross validation may lead to over optimistic performances with ANN. Although the generalization properties of the transfer function can be assessed by jack-knifing cross-validation in *PaleoNet*, K-fold cross-validation is also implemented for more robust evaluation of the predictive capacities. Details regarding neural network modelling, training parameter adjustment and cross-validation methods are provided in the *PaleoNet* users'guide (http://www.cen.ulaval.ca/paleo/).

The pruning option of *PaleoNet* for improved models

Although various types of algorithms have been proposed to model the current relationship between the organism and their environment, few of them offer the possibility to evaluate the relative contribution of each organism in the calibration. Generally, transfer functions are based on all organisms found in sufficient number in the calibration data-set assuming that each of these contributes equally to the model performance. However, it has been shown, at least for diatoms, that the contribution of species in several transfer functions was unequal (Racca et al. 2001, 2003, 2004) and that the omission of some numerically less important or non-contributing species in the calibration can improve the robustness and/or the performance of the models. Although more work is needed, it appears necessary to evaluate the relative contribution of the organisms in the transfer functions to exclude those that do not contribute to the empirical predictive performance of the models. The pruning option (HVS, Yacoub and Bennani 1997) of *PaleoNet* makes this possible. After a training simulation, the users will have the possibility to remove inputs (species) according to their relative contribution to the model performance, analogously to backward elimination in multiple regression analysis. In the HVS method, the contribution of a species i in the transfer function is evaluated using a heuristic: If x_i is an input neuron, w_{ij} is the weight on the connexion from this neuron to the neuron j of the hidden layer and w_{jo} is the weight on the connexion from neuron j to the output, the importance of input x_i can be estimated by:

$$\sum_{j\in H}\left(|w_{ij}| \Big/ \sum_{i\in I}(|w_{ij}|)\right) * w_{jo} \quad (1)$$

Where H is the hidden layer and I the input layer. The denominators:

$$\sum_{i\in I}(|w_{ij}|) \quad (2)$$

are used as a normalizing factor. This heuristic means that a species is more important if the path from this species to the output neuron has heavy weights. This method does not need the calculation of derivatives and is said to be a zero-order method. The importance of each input is thus

rapidly calculated and the less relevant can be suppressed. The neural network is then retrained. To show the effectiveness of the pruning algorithm of Yacoub and Bennani (1997), we applied it on various data-sets recently published in the literature. Even though the ANN-based transfer functions do not systematically show the lowest RMSEP, better ANN models were systematically obtained when the number of species was reduced (Table 1). In all cases considered here, the removal of non-contributing species improved the predictive capacity of the models.

Applying artificial neural network transfer functions with *PaleoNet*

Validated (pruned or not) transfer functions can be applied on fossil data-sets to produce environmental reconstructions. The "paleoenvironmental reconstruction" unit of *PaleoNet* is intended for the application of ANN-based transfer functions on sedimentary fossil data for which we desire quantitative inferences. *PaleoNet* will automatically adjust modern and fossil data files so that there is correspondence between them. A report of correspondence is then produced (i.e. species common to both data files; fossil species absent in the models but present in the sedimentary sequence, etc.). This type of report is very useful for the pruning step done during the development of the models.

Conclusions

This note briefly presented new software intended for the development, the evaluation and the application of ANN transfer functions in paleoecology. *PaleoNet* was developed with the aim of offering researchers in the paleosciences not only an additional but also a complementary method to those already available. The application of *PaleoNet* on a large range of calibration data-sets

Table 1 Summary of root mean square error of prediction of some recent organism–environment transfer functions published in the literature

Authors	Proxy	Variable	RMSEP (jack-knife)						
			WA	WA-PLS (comp)	ML	MAT	ANN	ANN with pruning	% taxa removed
Philibert and Prairie (2002)	Diatoms	pH	0.444	xxx	0.546	0.517	0.491	**0.444**	80
Philibert and Prairie (2002)	Diatoms	TP	0.519	xxx	0.589	0.477	0.470	**0.394**	60
Philibert and Prairie (2002)	Diatoms	DOC	0.425	xxx	0.539	0.430	0.412	**0.369**	80
Köster et al. (2004)	Diatoms	pH	0.297	**0.266 (3)**	0.308	0.343	0.319	0.282	20
Köster et al. (2004)	Diatoms	TP	0.242	**0.227 (2)**	0.265	0.283	0.265	0.242	60
Walker et al. (1991)	Chironomids	Temp	2.269	1.895 (2)	1.790	2.190	1.920	**1.200**	50
H. J. B. Birks (unpublished)	Pollens	Temp	1.184	**0.91(5)**	1.396	0.912	1.010	0.970	50
Gregory-Eaves et al. (1999)	Diatoms	Depth	0.324	0.312 (2)	0.344	0.303	0.319	**0.265**	80
Fallu et al. (2002)	Diatoms	Alkalinity	**0.264**	xxx	0.290	0.265	0.317	0.310	80
Fallu et al. (2002)	Diatoms	Color	0.236	**0.221 (2)**	0.262	0.224	0.256	0.233	60
Fallu and Pienitz (1999)	Diatoms	DOC	0.141	0.101 (3)	0.151	0.144	0.094	**0.072**	80
Larocque et al. (2001)	Chironomids	Temp	1.424	1.176 (5)	1.454	1.207	1.115	**1.033**	60
Bigler and Hall (2002)	Chironomids	Temp	1.194	1.091 (2)	1.248	1.322	1.029	**0.864**	60

WA, Weighted averaging regression/calibration; WA-PLS, Weighted averaging partial least square regression; ML, Maximum likelihood calibration; MAT, Modern analog technique; ANN, Backpropagation of artificial neural networks; TP, total phosphorus; DOC, dissolved organic carbon; Temp, temperature; Cond, conductivity; (comp), number of components

with different characteristics will enable researchers to determine where the backpropagation of ANNs is more appropriate than the regression/calibration in weighted averaging and vice versa. The use of *PaleoNet* will also allow the paleoecological scientific community to explore more deeply the important question of the relative numerical contribution of the organisms to the transfer functions.

Acknowledgements We would like to thanks John Birks, Christian Bigler, Marie-Andrée Fallu, Irene Gregory-Eaves, Dörte Köster, Isabelle Larocque, Aline Philibert and Ian Walker for their permission to use their calibration data set. PaleoNet was developed at the University of Nouvelle Calédonie. The development of this software has benefitted from support obtained through the NSERC-CRO project "Late Pleistocene Paleoclimate of Eastern Beringia" awarded to Les Cwynar, from NSERC operating grants awarded to R. Pienitz and Y.T. Prairie and from the Conseil National de la Recherche grants awarded to R. Racca.

References

Bigler C, Hall RI (2002) Diatoms as indicators of climatic and limnological change in Swedish Lapland: a 100-lake calibration set and its validation for paleoecological reconstructions. J Paleolimnol 27:97–115

Birks HJB, Line JM, Juggins S, Stevenson AC, ter Braak CJF (1990) Diatoms and pH reconstructions. Phil Trans Roy Soc Lond B 327:263–278

Fallu MA, Pienitz R (1999) Diatomées lacustres de Jamésie-Hudsonie (Québec) et modèle de reconstitution des concentration de carbone organique dissous. Ecoscience 6:603–620

Fallu MA, Allaire N, Pienitz R (2002) Distribution of freshwater diatoms in 64 labrador (Canada) lakes: species–environment relationships along latitudinal gradients and reconstruction models for water colour and alkalinity. Can J Fish Aquat Sci 59:329–349

Gregory-Eaves I, Smol JP, Finney BP, Edwards ME (1999) Diatom-based transfer functions for inferring past climatic and environmental changes in Alaska, USA. Arct Antarct Alp Res 31:353–365

Imbrie J, Kipp NG (1971) A new micropaleontological method for quantitative paleoclimatology: application to a late Pleistocene Carribean core. In: Turekian KK (ed) The late cenozoic glacial ages. Yale University Press, New Haven and London, pp 71–181

Juggins S, ter Braak CJF (1993) CALIBRATE – a program for species-environment calibration by weighted averaging partial least squares regression. Department of Geography, University of Newcastle, Newcastle upon Tyne, UK

Juggins S, ter Braak CJF (1995) WAPLS. Unpublished computer program, version 1.0, Department of Geography, University of Newcastle, Newcastle upon Tyne, UK

Juggins S (2004) C^2 User guide. Software for ecological and paleoecological data analysis and visualisation. University of Newcastle, Newcastle upon Tyne, UK, 69 pp\

Köster D, Racca JMJ, Pienitz R (2004) Diatom-based inference models and reconstructions revisited: methods and transformations. J Paleolimnol 32:233–245

Kucera M, Weinelt M, Kiefer T, Pflaumann U, Hayes A, Chen MT, Mix AC, Barrows TT, Cortijo E, Duprat J, Juggins S, Waelbroeck C (2005) Reconstruction of sea-surface temperatures from assemblages of planktonic foraminifera: multi-tecnique approach based on geographically constrained calibration data sets and its application to glacial Atlantic and Pacific Oceans. Quaternary Sci Rev 24:951–998

Larocque I, Hall RI, Grahn E (2001) Chironomids as indicators of climate change: a 100-lake training set from a subarctic region of northern Sweden (Lapland). J Paleolimnol 26:307–322

Malmgren BA, Nordlund U (1997) Application of artificial neural networks to palaoceanographic data. Palaeogeogr Palaeoclimatol Palaeoecol 136:359–373

Malmgren BA, Kucera M, Nyberg J, Waelbroeck C (2001) Comparison of statistical and neural network techniques fro estimating past sea-surface temperatures from planktonic foraminifer census data. Paleoceanography 16:520–530

Philibert A, Prairie YT (2002) Diatom-based transfer functions for western Quebec lakes (Abitibi and Haute Maurice): the possible role of epilimnetic CO2 concentration in influencing diatom assemblages. J Paleolimnol 27:465–480

Racca JMJ, Philibert A, Racca R, Prairie YT (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 Paleolimnol 26:411–422

Racca JMJ, Wild M, Birks HJB, Prairie YT (2003) Separating wheat from chaff: diatom taxon selection using an artificial neural network pruning algorithm. J Paleolimnol 29:123–133

Racca JMJ, Gregory-Eaves I, Pienitz R, Prairie YT (2004) Tailoring palaeolimnological diatom-based transfer functions. Can J Fish Aquat Sci 61:2440–2454

Reed R (1993) Pruning algorithms – a survey. IEEE Trans Neural Networks 4:740–747

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representation by back-propagating errors. Nature 323:533–536

Telford RJ, Birks HJB (2005) The secret assumption of transfer functions: a problems with spatial autocorrelation in evaluating model performance. Quaternary Sci Rev 24:2173–2179

ter Braak CJF, van Dam H (1989) Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia 178:209–223

ter Braak CJF, Juggins S (1993) Weighted averaging partial least squares regression (WA-PLS): an