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
© Springer Science+Business Media B.V. 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

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 paleoecology 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\textsubscript{i} is an input neuron, w\textsubscript{ij} is the weight on the connexion from this neuron to the neuron j of the hidden layer and w\textsubscript{jo} is the weight on the connexion from neuron j to the output, the importance of input x\textsubscript{i} can be estimated by:

\[
\sum_{j \in H} \left( \frac{\left| w_{ij} \right|}{\sum_{i \in I} \left( \left| w_{ij} \right| \right)} \right) \ast w_{jo} \quad (1)
\]

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

\[
\sum_{i \in I} \left( \left| w_{ij} \right| \right) \quad (2)
\]

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>
<thead>
<tr>
<th>Authors</th>
<th>Proxy</th>
<th>Variable</th>
<th>RMSEP (jack-knife)</th>
<th>% taxa removed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>WA</td>
<td>WA-PLS (comp)</td>
</tr>
<tr>
<td>Philibert and Prairie (2002)</td>
<td>Diatoms</td>
<td>pH</td>
<td>0.444</td>
<td>xxx</td>
</tr>
<tr>
<td>Philibert and Prairie (2002)</td>
<td>Diatoms</td>
<td>TP</td>
<td>0.519</td>
<td>xxx</td>
</tr>
<tr>
<td>Philibert and Prairie (2002)</td>
<td>Diatoms</td>
<td>DOC</td>
<td>0.425</td>
<td>xxx</td>
</tr>
<tr>
<td>Köster et al. (2004)</td>
<td>Diatoms</td>
<td>pH</td>
<td>0.297</td>
<td><strong>0.266 (3)</strong></td>
</tr>
<tr>
<td>Köster et al. (2004)</td>
<td>Diatoms</td>
<td>TP</td>
<td>0.242</td>
<td><strong>0.227 (2)</strong></td>
</tr>
<tr>
<td>Walker et al. (1991)</td>
<td>Chironomids</td>
<td>Temp</td>
<td>2.269</td>
<td>1.895 (2)</td>
</tr>
<tr>
<td>H. J. B. Birks (unpublished)</td>
<td>Polens</td>
<td>Temp</td>
<td>1.184</td>
<td><strong>0.91 (5)</strong></td>
</tr>
<tr>
<td>Gregory-Eaves et al. (1999)</td>
<td>Diatoms</td>
<td>Depth</td>
<td>0.324</td>
<td>0.312 (2)</td>
</tr>
<tr>
<td>Fallu et al. (2002)</td>
<td>Diatoms</td>
<td>Alkalinity</td>
<td><strong>0.264</strong></td>
<td>xxx</td>
</tr>
<tr>
<td>Fallu et al. (2002)</td>
<td>Diatoms</td>
<td>Color</td>
<td>0.236</td>
<td><strong>0.221 (2)</strong></td>
</tr>
<tr>
<td>Fallu and Pienitz (1999)</td>
<td>Diatoms</td>
<td>DOC</td>
<td>0.141</td>
<td>0.101 (3)</td>
</tr>
<tr>
<td>Laroque et al. (2001)</td>
<td>Chironomids</td>
<td>Temp</td>
<td>1.424</td>
<td>1.176 (5)</td>
</tr>
<tr>
<td>Bigler and Hall (2002)</td>
<td>Chironomids</td>
<td>Temp</td>
<td>1.194</td>
<td>1.091 (2)</td>
</tr>
</tbody>
</table>

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 Phillibert 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 benefited from support obtained through the NSERC-CRO project “Late Pleistocene Paleoclimate of Eastern Beringia” awarded to Les Cywany, 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


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


