Artificial neural networks as a tool for predicting fish community composition in rivers

M. Scardi[a*], S. Lek[b], P. Lim[c], P. Di Dato[a], T. Oberdorff[d]

[a]Department of Zoology, University of Bari, Via Orabona 4, 70125 Bari

[b]CNRS - UMR 5576, CESAC, Université Paul Sabatier, Bât. IVR3, 118 Route de Narbonne, F-31062 Toulouse Cedex, France

[c]ENSAT, Equipe Environnement Aquatique & Aquaculture, Avenue de l'Agrobiopole, BP 107, 31326 Castanet Tolosan, France

[d]Muséum National d'Histoire Naturelle, Laboratoire d'Ichtyologie Générale et Appliquée, 43 Rue Cuvier, F-75231 Paris Cedex 05, France

Fish community composition has been successfully modeled by means of artificial neural networks (ANNs), using environmental variables as predictors. Two applications of this technique are presented, based on different levels of integration of the information about fish community composition. Local species richness was modeled in the Garonne river (France), using only 3 environmental predictive variables, whereas presence or absence of 30 fish species was modeled in north-eastern Italian rivers using 26 environmental predictive variables. The results were very good in both cases: the ANN models were able to explain 82% of the variance in local species richness and to correctly predict presence or absence of fish species in more than 85% of the cases. These results pointed out that ANNs can play a very important role in the study of ecological communities.

1. INTRODUCTION

Artificial neural networks (ANNs) are powerful tools for empirical modeling, that can be applied to a broad spectrum of problems, even when the underlying causal relationships are poorly understood or completely unknown. Moreover, any finite-dimensional vector function on a compact set can be approximated to arbitrary precision by multilayer feedforward ANNs, provided that enough data and computing resources are available [1].

As ANNs can model complex nonlinear relationships, they have proved to be very useful in ecological research, where these conditions are very common [2]. In particular, species distribution with respect to complex environmental gradients is a typical example of multiple nonlinear biotic response. Its study has played an important role in modern ecology, since it is deeply related to the development of the concept of ecological community.

*Email: mscardi@mclink.it

Given a data set large enough to allow for effective ANN training (i.e. for model calibration), many ecological variables can be accurately estimated on the basis of both biotic and abiotic predictive variables [3]. Even the prediction of ecological community structure, which is a very challenging problem for conventional models [4], could be successfully solved by using ANNs, as previous attempts have already pointed out [5–7, etc.].

In this paper the above said problem will be dealt with in relation to river ecosystems at two very different levels of integration, by predicting: (a) local fish species richness, i.e. the total number of species at a given site; (b) fish community composition, i.e. presence or absence of fish species at a given site.

The former goal is probably simpler to attain, but it is also the most promising in terms of generalization. The latter one, although not trivial to achieve, might play an important role in ecological research, as it could provide significant insights into the inner mechanics of community ecology and overcome the limitation of conventional linear approaches to ecological data analysis and modeling.

2. MATERIALS AND METHODS

2.1. Data collection

Data from 207 sampling sites have been used to model fish local species richness (LSR) in the Garonne river (France). LSR data were collected between 1986 and 1996 by using electrofishing, during low-flow periods. At each studied site, three environmental variables were recorded: distance from source (DFS), elevation (ELE) and catchment area (CAA). CAA was measured with a digital planimeter on a 1/500000 scale map of the Garonne river basin, whereas the other variables (DFS and ELE) were collected from 1/25000 scale maps. A second independent data set, including data for the same variables from 72 sites, was used to test the model. It has been collected according to the same sampling procedures from 1985 to 1995.

Data from 176 sampling sites in the rivers of the province of Vicenza (Italy) have been used to predict fish community structure on the basis of 26 environmental variables, namely: elevation (m); falls, small falls, rapids, runs, pools and riffles (surface, %); runs, pools and riffles (depth, cm); maximum, minimum and average width (m); homogeneity (score, 0-5); boulders, rocks and pebbles, gravel, sand, silt and clay (surface, %); stream velocity (0-5); flow ($dm^3\ s^{-1}$); vegetation covering (surface, %); shadow (0-5); anthropic disturbance (0-5); pH; conductivity (μS). All the environmental data were scaled into a [0,1] interval before ANN training. Information about fish community was based on binary (i.e. presence or absence) data for 30 species (see species names in tab. 1). Sampling was carried out by electrofishing from 1987 to 1994.

2.2. Modeling methods

A three layer feedforward ANN was used to model LSR. It had a 3-3-1 architecture, i.e. an input layer of three neurons (one for each environmental variable), a hidden layer with three neurons and an output layer with a single neuron (i.e. LSR). Sigmoid activation functions have been used in all the nodes of the hidden and output layers of the ANN. Learning rate and momentum, which were initially fixed respectively at 0.01 and 0.95, have been modified during the training according to the output error. The initial weights for the links between neurons were randomly chosen within a [-0.3,0.3] interval.

The error backpropagation algorithm [8] was used to adjust these weights during the training procedure. Modeling was carried out in three steps: (1) the first data set (n=207) was randomly divided into two data subsets and the first subset, i.e. 75% of whole set (n=155), was used to train the ANN model for 500 epochs; (2) the second subset, i.e. 25% of the whole set (n=52), was used to test the previous model; (3) further testing of the ANN model was performed using the second independent data set (n=72) in order to determine the predictive quality of the model.

The same ANN architecture, i.e. a three layer feedforward ANN, was also used to model fish community composition, i.e. presence or absence fish species. In this case, however, the ANN was much more complex, as it had a 26-40-30 structure. The number of nodes in the hidden layer was defined after empirical tests, that showed that 40 hidden nodes provided the best performance when compared to other structures with a number of hidden nodes in the $[\frac{n}{2},2n]$ range, where n is the number of input nodes.

The training procedure has been slightly more complex for this ANN than for the previous one. As in the previous case, an "early stopping" strategy was used, but the number of epochs was not *a priori* defined. In fact, all the available data were divided into three sets. Half of the available patterns (n=88) was randomly selected as training set. A second set (25% of the available patterns, n=44) was used to compute the mean square error (MSE) of the ANN output after each training epoch: as soon as the MSE started to increase, the training was stopped. Finally, a third data set (25% of the available patterns, n=44) was only used to test the performance of the ANN after completion of the training phase.

The learning constant was set to 0.9 and the momentum term to 0.1 and these values were not modified during training. In order to avoid overtraining, a small amount of gaussian noise (μ=0, σ=0.01) was added to the ANN input patterns (thus generating a virtually infinite number of slightly different input patterns). Moreover, only a random subset of the training set was submitted to the ANN at each epoch (i.e. 44 patterns out of 88) .

3. RESULTS

The 3-3-1 ANN model of local fish species richness (LSR) performed very well, as the determination coefficient for the regression between observed values and ANN outputs obtained after training was R^2=0.874 (Figure 1). When computed only for the test subset in the first data set the determination coefficient was R^2=0.817, whereas it was somewhat smaller (R^2=0.676) when the second independent data set was used as a test set.

The 26-40-30 ANN model of fish community structure, in spite of its inner complexity, provided accurate estimates of species presence or absence. After completion of the training procedure, all the ANN outputs, which were continuous because of the sigmoid activation function in the output layer, were converted back to a binary form by a threshold function that was set at 0.5. The percentages of matches, i.e. of exact predictions of species presence or absence that were obtained with respect to the test set, are shown in Table 1.

As the test set was completely independent of the training and validation sets, the accuracy of the ANN model was remarkable: in fact, the average percentage of matches

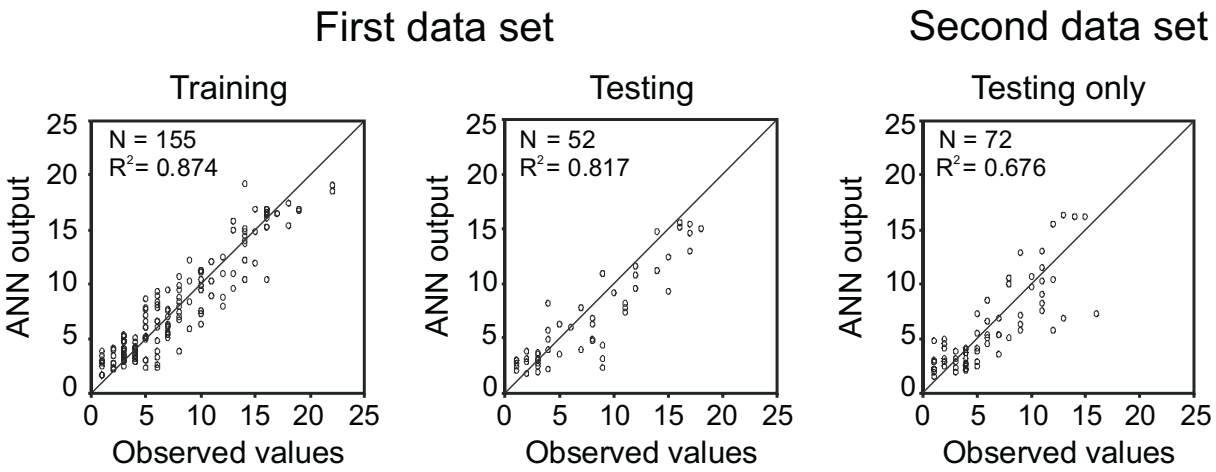

Figure 1. Garonne river local fish species richness: artificial neural network output vs. observed values.

was 85.6%. When predictions for single species were taken into account, in 14 out of 30 cases the percentage of matches was greater than 90%, whereas only one species (*Cottus gobio*) was almost unpredictable, as only 52.3% of matches between its field records and ANN output were obtained (obviously random predictions would have provided 50% of matches).

It is important to notice that the low predictability of this species was ecologically sound, as *C. gobio* can be found in two completely separated river zones and its presence or absence depends on water quality descriptors that were not included in the set of environmental predictive variables of this ANN model.

However, the level of predictability of a given species can be affected not only by the relevance of the environmental predictive variables, but also by the role the species plays in the ecosystem. For instance, species whose level of predictability is low are probably more involved in complex biotic relationships (e.g. interspecific competition) than species whose presence or absence can be accurately predicted by the ANN model on the basis of the given set of abiotic environmental variables. Therefore, the level of predictability provided by the ANN model can highlight interesting ecological properties of the modeled species.

When the whole community structure was taken into account instead of the distributions of single species, the ANN predictions still proved to be accurate. In order to compare real and predicted community structure, two matrices of similarity between fish community samples were computed, using real data from the test set and data predicted by the the ANN model on the basis of environmental information. The similarity between real and predicted fish community compositions was computed according to the Rogers & Tanimoto index [9]:

$$S_{ij} = \frac{a+d}{a+2b+2c+d} \tag{1}$$

where a and d are respectively the number of species whose presence and absence matched

Table 1
Province of Vicenza fish communities: percentage of matches in estimates of species presence or absence (test set only).

Species name	matches	Species name	matches
Abramis brama	97.7%	*Perca fluviatilis*	97.7%
Alburnus alburnus alborella	77.3%	*Lepomis gibbosus*	90.9%
Anguilla anguilla	75.0%	*Micropterus salmoides*	93.2%
Barbus plebejus	79.5%	*Ictalurus melas*	97.7%
Carassius carassius	93.2%	*Phoxinus phoxinus*	81.8%
Cyprinus carpio	100.0%	*Scardinius erythrophthalmus*	88.6%
Leuciscus cephalus	75.0%	*Cottus gobio*	52.3%
Cobitis taenia	72.7%	*Gasterosteus aculeatus*	90.9%
Gambusia holbrooki	93.2%	*Thymallus thymallus*	90.9%
Padogobius martensii	72.7%	*Tinca tinca*	84.1%
Gobio gobio	93.2%	*Rutilus erythrophthalmus*	72.7%
Lampetra planeri	95.5%	*Salmo (trutta) trutta*	88.6%
Chondrostoma genei	86.4%	*Oncorhynchus mykiss*	84.1%
Esox lucius	79.5%	*Salmo (trutta) marmoratus*	100.0%
Orsinigobius punctatissimus	72.7%	*Leuciscus souffia*	90.9%

the ANN predictions, and b and c are the number of species which were present but not predicted and vice versa.

These similarity matrices were compared by using the Mantel permutation test [10], which allowed to reject the null hypothesis of no relationships between them [R=0.50778, P(R)=1.000, n=100000]. As the unit probability level suggests, the observed value of the standardized Mantel statistics R was greater than all those that were obtained after 100000 random permutations of one of the similarity matrices. This obviously implies that the fish community composition was not only non-independent in observed and predicted data sets, but even very similar. This evidence is particularly significant, because the Rogers & Tanimoto similarity index allocates more importance (i.e. a double weight) to discrepancies than it does to agreements between observed and predicted data.

4. CONCLUSIONS

ANNs proved again to be a very important tool in modern ecological research, as they can be effectively used to predict community composition on the basis of environmental variables. ANNs can outperform most conventional modeling and data analysis techniques because they are inherently more suited to deal with nonlinear relationships, which are the rule rather than the exception in ecology.

They can be applied to a wide spectrum of ecological problems [3], providing accurate estimates of synthetic descriptors (e.g. species richness) as well as complex information like species composition of ecological communities.

Even though a proper selection of relevant predictive variables is always of paramount importance in ecological modeling and no model can work properly if enough information

about the underlying processes is not available, ANN models can effectively cope with ill-defined problems. In fact, ANN models can adjust their inner structure in order to isolate ineffective predictive variables (i.e. unnecessary ANN inputs), provided that their training processes have been optimized to attain optimal accuracy and generalization capabilities.

5. ACKNOWLEDGEMENTS

This work was funded by the 5^{th} EC Framework programme (project PAEQANN, EVK1-CT1999-00026).

The data set from the Garonne river which has been used for ANN testing only was provided by the Conseil Supérieur de la Pêche (France). All the data from the rivers of the province of Vicenza (Italy) have been collected by Aquaprogram s.r.l. and are available on the Web at http://www.provincia.vicenza.it/mappaitt/index.htm.

REFERENCES

1. K. Hornik, M. Stinchcombe and H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359-366.
2. S. Lek, M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga and S. Aulanier, Application of neural networks to modelling nonlinear relationships in ecology, Ecological Modelling 90 (1996) 39-52.
3. S. Lek and J-F. Guégan (eds.), Artificial Neuronal Networks, Applications to Ecology and Evolution, Springer-Verlag, Berlin-Heidelberg, Germany, 2000.
4. M.K. Joy and R.G. Death, Development and application of a predictive model of riverine fish community assemblages in the Taranaki region of the Horth Island, New Zealand, New Zealand Journal of Marine and Freshwater Research 34(2) (2000) 241-252.
5. P. Boët and T. Fuhs, Predicting presence of fish species in the Seine river basin using artificial neural networks, in: S. Lek and J-F. Guégan (eds.), Artificial Neuronal Networks, Applications to Ecology and Evolution, Springer-Verlag, Berlin-Heidelberg, Germany, (2000) 131-142.
6. T.-S. Chon, Y.S. Park, K.H. Moon and E.Y Cha, Patternizing communities by using an artificial neural network, Ecological Modelling 90 (1996) 69-78.
7. F. Recknagel, ANNA - Artificial neural network model for predicting species abundance and succession of blue-green algae, Hydrobiologia 349 (1997) 47-57.
8. D.E. Rumelhart, G.E. Hinton and R.J. Williams, Learning representations by back-propagating error, Nature 323 (1986) 533-536.
9. D.J. Rogers and T.T. Tanimoto, A computer program for classifying plants, Science (Wash. D.C.) 132 (1960) 1115-1118.
10. N. Mantel, The detection of desease clustering and a generalized regression approach, Cancer Research 27 (1967) 209-220.