Limnol. Oceanogr., 44(5), 1999, 1293–1303

Predicting fish distribution in a mesotrophic lake by hydroacoustic survey and artificial neural networks

Sébastien Brosse and Sovan Lek
CNRS, UMR 5576 CESAC—Université Paul Sabatier, 118 Route de Narbonne 31062 Toulouse cedex, France

Francis Dauba
ENSAT, Laboratoire d'Ingénierie agronomique, Equipe Environnement Aquatique et Aquaculture, avenue de l'Agrobiopole, BP 107, 31326 Castanet-Tolosane cedex, France

Abstract

The present work describes the development and validation of Artificial Neural Networks (ANN) by comparison with classical and more advanced parametric and nonparametric statistical modeling methods such as Multiple Regression (MR), Generalized Additive Models (GAM), and Alternating Conditional Expectations (ACE) to estimate spatial distribution of fish in a mesotrophic reservoir. The modeling approaches were developed and tested using 60 hydroacoustic transects covering the whole lake. Each transect was divided into 100-m-long sections, constituting a total of 732 sampling units. For each of them, the relationships between topology, chemical characteristics, and fish abundance were studied. The models had six independent topological (i.e., depth, distance from the bank, slope, and stratum) and chemical (i.e., temperature and dissolved oxygen) variables and one dependent output variable (fish density, FD). The data matrix was divided into two parts. The first contained units where FD was nonnil (i.e., 399 sampling units), and the second contained only cases without fish (i.e., 333 sampling units). Model training and testing procedures were run on the first submatrix after log(FD + 1) transformation. As linear MR results were not satisfactory ($r^2 = 0.42$ in the training set, and $r^2 = 0.51$ in the testing set) compared with ANN ($r^2 = 0.81$ in the training set, and $r^2 = 0.77$ in the testing set), we tried nonlinear transformations of the variables such as logarithmic, lowess (for the GAM), and an optimal nonlinear transformation using the SAS Transreg procedure (for the ACE model), but the determination coefficients remained clearly lower than those obtained using ANN ($r^2 = 0.60$ in the training set for ACE, and $r^2 = 0.66$ in the training set for GAM). The results of a second test on the nil submatrix stressed that, compared with other statistical techniques, ANN and, to a certain extent, GAM models were able to clearly define the potential FDs in samples where no fish were actually found. The model showed, on the basis of the topological and chemical variables taken into account, that the predicted potential FDs in the surface stratum are higher than in the underlying stratum. Finally, on the basis of the sensitivity analyses performed on the ANN and GAM results, we established relationships between FDs and the six environmental variables. Our results exhibit a clear summer habitat preferendum, the fish (predominantly roach) being located mainly in the surface stratum, in the warm shallow littoral areas. These observations led us to discuss the ecological significance of such a fish distribution, which may be due to a trade-off between feeding, predation avoidance, and endogenous fish requirements.

Hydroacoustics is a well-established and recognized technique for the assessment and management of aquatic resources (Thorne 1983). In freshwater habitats, echograms have been used to describe vertical (Eggers 1978; Matthews et al. 1985; Hamrin 1986), spatial (O'Brien et al. 1984; Burczynski et al. 1987; Hruska 1989), and temporal (Bohl 1980; Baker and Paulson 1983; Imbrock et al. 1996) patterns in fish distribution. Nevertheless, very few studies have simultaneously taken into account the relationships between fish spatial distribution and several environmental variables. Consequently, the goal of our work was to model the spatial distribution of lake fish populations according to descriptors of the physical and chemical characteristics and the topological environment, allowing a predictive model of fish distribution to be set up using easily measurable variables. Conventional techniques, based notably on MR, are capable of solving many problems, but they sometimes show serious shortcomings (James and McCulloch 1990). This difficulty occurs because relationships between variables in environmental sciences are often nonlinear, while the methods used are intrinsically linear. Nonlinear transformations of the dependent and/or independent variables (e.g., logarithmic, power, or exponential functions) can improve the results but often not entirely satisfactorily (Lek et al. 1996*b*). To address this deficiency, several nonparametric modeling methods have been set up such as GAM (Hastie and Tibshirani 1990), ACE (Young 1981), and, more recently, ANN. ANN, with the error backpropagation procedure, provides a nonlinear alternative to linear regression, particularly with nonlinear relations (Rumelhart et al. 1986). Recently, certain applications of ANN in aquatic ecology have been published, e.g., the use of ANN for modeling stream hydrobiological and ecological responses to climate changes (Poff et al. 1996), prediction of phytoplankton production (Scardi 1996), identification of the major goals of underwater acoustics (Cas-

Acknowledgments
We thank the two anonymous reviewers for constructive comments on an earlier version of this paper. We are grateful to P. Winterton for correcting the English version. We thank P. Sarda and I. Dimopoulos for providing helpful criticism during the preparation of this manuscript. This research was supported in part by a doctoral grant (S.B.) provided by the French electricity agency.

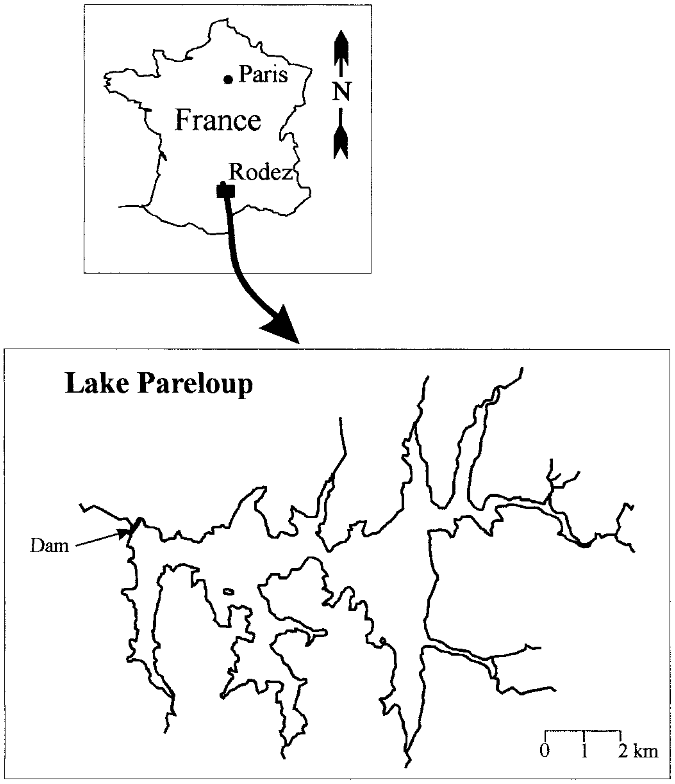

Fig. 1. Map of France showing location of Lake Pareloup and a representation of the reservoir.

selman et al. 1994), prediction of global fish species richness (Guégan et al. 1998), prediction of density of brown trout redds (Lek et al. 1996*b*), and prediction of density and biomass of various fish species (Baran et al. 1996; Lek et al. 1996*a;* Mastrorillo et al. 1997).

In this paper, the capacities of ANN and nonparametric regression methods (i.e., GAM and ACE) are compared for an MR-type problem. In addition to its primary goal, i.e., modeling the relationships between the ecological variables of the environment and the fish distribution in a reservoir, the present work aims to quantify the influence of six environmental variables on fish spatial distribution using ANN and GAM, leading to the proposal of hypotheses about its ecological significance.

Materials and methods

Study site and sampling—Lake Pareloup was selected as the study site because of its structural heterogeneity in the form of numerous bays with a wide range of topographical characteristics (Fig. 1). This reservoir is located in the southwest of France, near the city of Rodez. It covers an area of 1,350 ha, for a volume of about 168×10^6 m^3. The maximum depth is 37 m, and the average depth is 12.5 m. Lake Pareloup is a warm monomictic lake, which is therefore subjected to a summer thermal stratification, with a low oxygen content below the thermocline (located at about 10-m depth from early June to mid-September) preventing the fish from colonizing deep water during this period.

The survey was carried out during thermal stratification (July), when the overwhelming majority of fish do not live at the bottom (Schultz 1988; Hruska 1989). Sixty transects covering the different parts of the lake were performed between 0900 and 1930 h to avoid miscalculations that could be caused by diurnal fish migrations in the early morning and in the evening between shore and open water (Hasler and Villemonte 1953; Bohl 1980; Imbrock et al. 1996). The transducer was mounted at the front of a 4-m-long boat that cruised at a speed of 4.5 knots. Acoustic data were collected with a Lowrance X-15MA echosounder, operating frequency 192 kHz, equipped with a vertical single-beam transducer towed 0.3 m below the surface and aimed straight down. Total beam angle was 20° measured at the −3 dB level. All pulses were transmitted to the 20° transducer element, and pulse duration was 0.2 ms. The signals were then amplified by the echosounder at 20 $\log_{10}$ (R) Time Varied Gain (TVG) and relayed to the chart recorder and the echointegrator. The sound level was set at a value such that fish below 60-mm total length (−60 dB) and other small targets were not detected.

Information provided by the echosounder was recorded on a portable computer using SIBELIUS software (Richeux et al. 1994), which vertically divides space into 50-cm strata and gives the number of fish detected for each 10-m echogram. The boat position (±10 m) from a Global Positioning System (GPS) and lake depth were memorized continuously. Finally, before and several times during the survey, the system was calibrated using a 32-mm-diameter copper sphere. The accuracy of the system approached ±1 dB or ±10% (Foote 1982; MacLennan and Simmonds 1992).

Each transect chart was divided into subsamples 100 m long and 5 m thick (1–6 m, 6–11 m), defining 732 sampling units. Shallow littoral areas (<2 m deep), the first meter below the surface, and also the last meter above the bottom were excluded from the transects because of the dark zone created by surface and bottom noise. For each of the sampling units, fish echotraces were numbered, and six environmental variables were considered. Mean depth (DEP) expressed in meters from the surface, mean slope of the bottom of the lake (SLO) expressed in percentage, and stratum sampled (STR) were provided by the echogram. DIS was the distance in meters between the middle of each unit and the closest bank. For STR, two strata 5 m thick were taken into account, i.e., 1–6 m and 6–11 m below the surface. For each sounded unit, oxygen concentration (milligram per liter) and water temperature (°C) were measured each meter from 1- to 11-m depth. OXY and TEM are the mean values of oxygen concentration and water temperature, respectively, for each 100-m-long, 5-m-thick unit.

Technique of ANN modeling—For ANN modeling, a multilayer feed-forward neural network was used. The processing elements in the network, called neurons, are arranged in a layered structure. The first layer, called the input layer, connects with the input variables. In our case, was comprised of six neurons corresponding to the six environmental variables. The last layer, called the output layer, connects to the output variable (Fig. 2). It was comprised of a single neuron corresponding to the value of the dependent variable to be

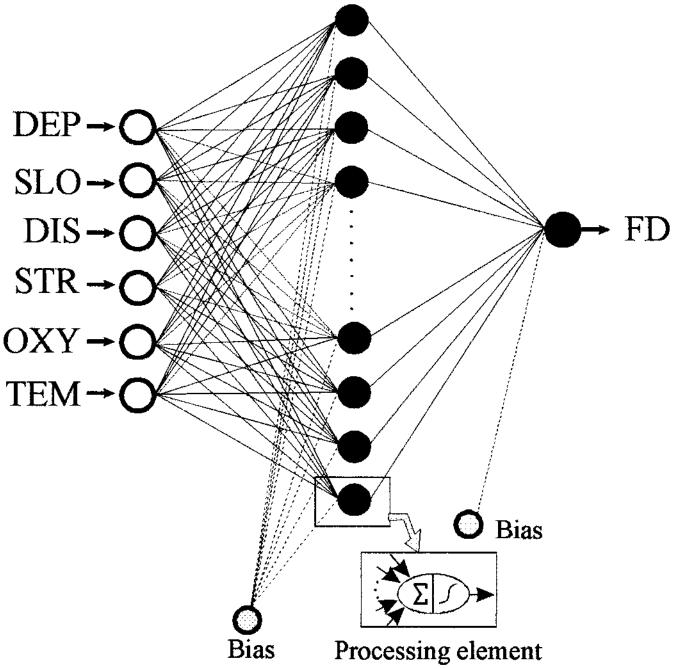

Fig. 2. Typical three-layered feed-forward artificial neural network. Six input neurons corresponding to 6 independent variables, 10 hidden layer neurons, and 1 output neuron for estimating FD. Connections between neurons are shown by solid lines; they are associated to synaptic weights that are adjusted during the training procedure. The bias neurons are also shown, and their input value is 1. The sigmoid activation function is used as a transfer function in the hidden and output layers.

predicted (FD). The layers between the input and output layers are called the hidden layers. There can be one or more hidden layers, and the number of neurons in each layer is an important parameter of the network. At some point during the training phase, the network passes through a configuration that gives the best generalization (Geman et al. 1992; Smith 1993). After this point, what the network learns amounts to overfitting, i.e., incapacity of the model to generalize. Thus, the network configuration is approached empirically by testing various possibilities (i.e., number of hidden neurons and number of iterations) and selecting the network configuration giving the best generalization without overfitting, i.e., the best compromise between bias and variance (Geman et al. 1992; Kohavi 1995). Each neuron is connected to all neurons of adjacent layers (neurons within a layer and in nonadjacent layers are not connected). Neurons receive and send signals through these connections. In feed-forward networks, signals are transmitted only in one direction: from input layer to output layer through hidden layers (no feed-back connections are permitted). Connections are given a weight that modulates the intensity of the signal they transmit. The weight plays an important role in propagation of the signals through the network. They establish a link between the input variables and their associated output variable and are said to contain the knowledge of the neural network about the problem–solution relation.

Training the network consists of using a training data set to adjust the connection weights in order to minimize the

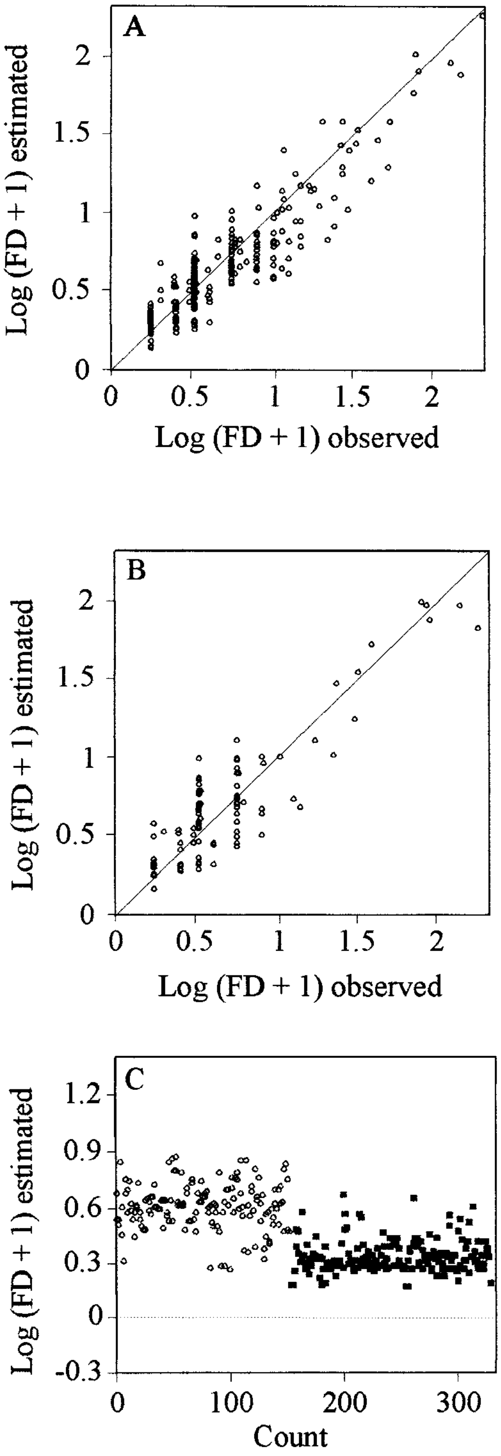

Fig. 3. ANN model predictions of FDs. (A, B) Scatterplots of predicted values vs. observed values in the training set (A) and in the testing set (B). The solid line indicates the perfect fit line (coordinates 1 : 1). *See text for the detail.* (C) Second test sorted for the two strata; open circles represent FD in the surface stratum, and black squares represent FD in the underlying stratum.

Table 1. Results of the ANN, linear MR, and GAM models on the five random training sets (two-thirds of SM1, i.e., 300 records) and five testing sets (the remaining one-third of records from SM1, i.e., 99). SD, standard deviation.

	ANN		MR		GAM	
	Training	Testing	Training	Testing	Training	Testing
Test 1	0.834	0.795	0.415	0.553	0.625	0.696
Test 2	0.815	0.785	0.399	0.574	0.688	0.520
Test 3	0.796	0.748	0.432	0.486	0.684	0.505
Test 4	0.769	0.706	0.454	0.422	0.657	0.589
Test 5	0.826	0.790	0.422	0.510	0.645	0.623
Mean	0.808	0.765	0.424	0.509	0.660	0.587
SD	0.026	0.038	0.020	0.060	0.024	0.051

error between observed and predicted values. This training was performed according to the backpropagation algorithm (Rumelhart et al. 1986). The connection weights, initially taken at random in the range of -0.3 to 0.3, are iteratively adjusted by a method of gradient descent based on the difference between the observed and expected outgoing signals. Many iterations are necessary to guarantee the convergence of predicted values toward their expectations without overfit. The computational program was realized in a Matlab® environment and computed with an Intel Pentium® processor.

Modeling was carried out after $\log(x + 1)$ transformation of FD, which was applied to avoid an undue influence of outliers on the determination coefficients (ter Braak and Looman 1995). Then, the whole data matrix (i.e., 732 records × six environmental variables) was divided in two submatrices, one (submatrix SM1) containing records with no zero values for FD (i.e., 399) and another (submatrix SM0) with the records where no fish were detected (i.e., 333). The quality of the assignment was judged by the hold-out procedure (Kohavi 1995) to determine recognition performance (training set) and prediction performance (testing sets). Modeling was carried out in three steps. (1) The model was trained after isolation, by random selection, of a training set from SM1 (SM1train: three-fourths of the records from SM1, i.e., 300). This step allows the performance of the ANN to be estimated. The determination coefficient between observed and predicted values was used to quantify the ability of the model to produce the right answer through the training procedure (recognition performance). (2) The model obtained during the training procedure was tested with the first test set (SM1test) constituted by the remaining one-fourth of the records from SM1 (i.e., 99). During the testing procedure, only the input variables were introduced in the network. This step allows the prediction capabilities of the network to be assessed. This operation (i.e., random selection of the SM1train and SM1test submatrix used in the two above-mentioned steps) was repeated five times, giving rise to ‘‘test1’’ to ‘‘test5’’ to study the stability of the FD predictions. Finally, (3) submatrix SM0 was used as a second test set. The model obtained with SM1train was tested with SM0. This step allowed us to estimate the potential FDs in the areas where no fish were detected.

Technique of MR modeling—For MR, calculations were done using SPSS® software (Norusis 1993). MR models were trained and tested using the same data sets as for ANN modeling (SM1train to establish the models and SM1test and SM0 for the first and the second tests of the models). First, only the dependent variable (FD) was $\log(x + 1)$ transformed, with an aim toward comparing the two methods (ANN and MR). Then, we tried to linearize the independent variables for which distribution was nonnormal (DIS, DEP, and SLO), using a log transformation for DIS and DEP and a $\log(x + 1)$ transformation for SLO. Aiming to improve the model’s performance, we also used GAM (Hastie and Tibshirani 1990). GAM models are a generalization of multiple linear regression and generalized linear models. They are nonparametric regression methods that model the dependent variable as an additive sum of unspecified functions of covariates. Least-squares and maximum likelihood methods used in multiple linear regression and generalized linear models are replaced by quasilikelihood methods that rely on local scatterplot smoothing methods. Here, we used the locally weighted smoother of Cleveland (1979), currently designed by ‘‘lowess’’ and called ‘‘loess’’ in the S-PLUS statistical computing language. The lowess smoother first computes a defined percentage of the nearest neighbors to the target point. A tricube kernel, centered at the target point, becomes zero at the furthest neighbor. The smoother at the target point is the fitted value from the locally weighted linear fit, with weights supplied by the kernel. One of the major advantages of this method is that it automatically shows the dependence of the response on each of the predictors. These results were compared to those obtained using ANN sensitivity analysis. GAM models were set up using S-PLUS® software. Finally, as a check, optimal nonlinear transformation was also tried on the training data set (SM1train), using the SAS Transreg procedure (SAS 1988). This procedure seeks an optimal transformation of variables, using a method of alternating least squares, to fit the data to a linear regression (Young 1981; Breiman and Friedman 1985; SAS 1988). Two complementary levels of optimal transformation were tried using a B-spline transformation. By default, a cubic polynomial transformation was used. To improve the transformation, knots (or break points) were specified. Each knot specifies a discontinuity in the *n*th derivative of the transformation function at the value of the knot. Knots can be repeated any number of times to decrease smoothness at the break points. In our case, we used two Transreg transfor-

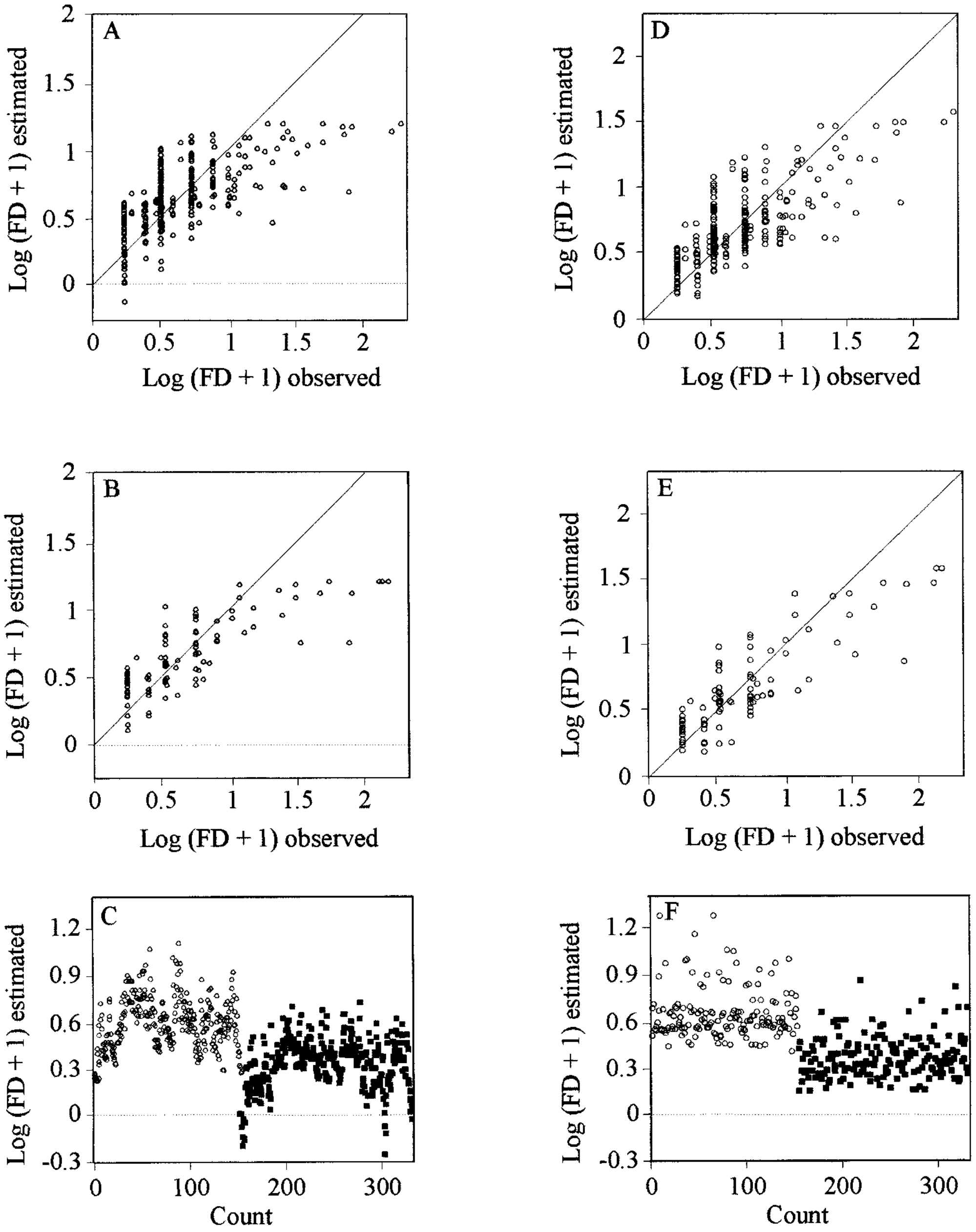

Fig. 4. Linear MR (A, B, C) and GAM (D, E, F) model predictions of FDs. (A, B, D, E) Scatterplots of predicted values vs. observed values in the training set (A, D) and in the testing set (B, E). The solid line indicates the perfect fit line (coordinates 1 : 1). *See text for details.* (C, F) Second test with linear MR (C) and GAM (F) sorted for the two strata; open circles represent FD in the surface stratum, and black squares represent FD in the underlying stratum.

mations, i.e., with one and two knots, aiming to obtain the best determination coefficient possible.

ANN sensitivity of independent variables—One disadvantage of ANN is its lack of explanation power. Some analyses, like MR, can identify the contribution each individual input makes on the output and can also give some measures of confidence about the estimated coefficients. By contrast, currently, there is no theoretical or practical way of accurately interpreting the weights attributed in ANN. For example, weights cannot be interpreted as a regression coefficient nor can difficulty be used to compute causal impacts or elasticities. Therefore, ANN models are generally better suited for forecasting or prediction than for policy analysis. But in ecology, it is necessary to know the impact of the explanatory variables. Some authors have proposed methods allowing the determination of the impact of the input variables (Garson 1991; Goh 1995; Lek et al. 1996*a*,*b*). In this work, Garson's algorithm and Lek's algorithm were used to determine the sensitivity of the environmental variables.

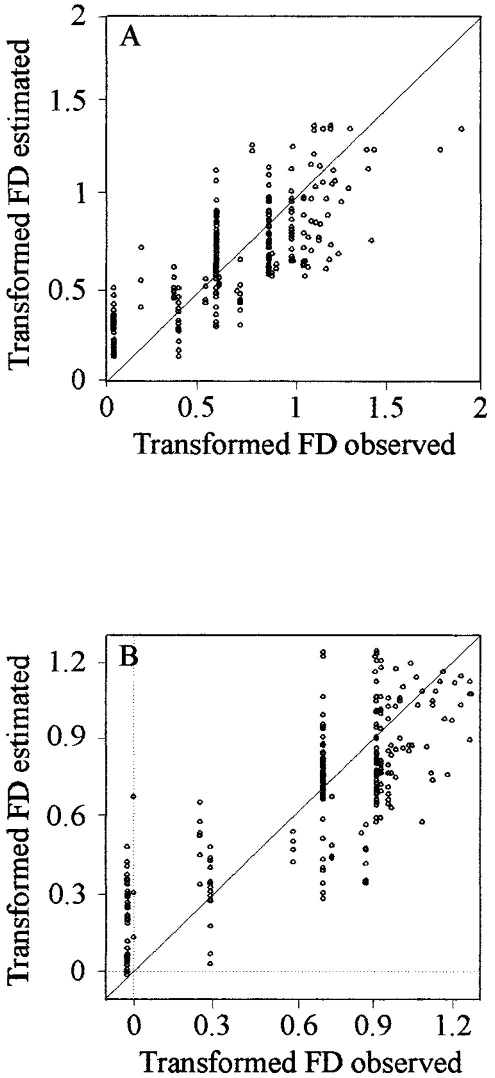

Fig. 5. Scatterplots of observed values vs. predicted values using ACE on the training set after the Transreg procedure with one (A) and two (B) knots. The solid line indicates the perfect fit line (coordinates 1 : 1).

Results

Large variations in FD were observed between samples, with a high coefficient of variation (345%). As a consequence, the distribution of the dependent variable shows a highly dissymmetric profile. This kind of distribution is frequent in many problems regarding plants or animals or with regard to spatial organization of populations. It can be qualified as skewed to the right (Jager and Looman 1995). Because normal distributions are more convenient to deal with, FDs that have a log-normal skewed distribution were transformed by taking their logarithm. After transformation, the coefficient of variation became <120%.

Artificial neural network—The ANN used was a three-layered (6 → 10 → 1) feed-forward network with bias. There were six input neurons to code the six independent variables. The hidden layer had 10 neurons, determined as the optimal configuration (networks with two hidden layers were not significantly better), giving the lowest error in the training and testing sets of data with minimal computing time (Lek et al. 1996*b,c*). The output neuron computes the value of the dependent variable (FD). We thus have a total of 81 parameters ((6 input neurons × 10 hidden neurons) + (10 hidden neurons × 1 output neuron) + 11 bias parameters). The calibration process has been illustrated by Lek et al. (1996*b,c*), where correlation coefficients of training and testing sets of data are plotted vs. the number of hidden neurons and number of iterations.

Figures 3A,B show the scatterplots between observed and predicted values of FDs by the ANN models after a learning procedure (Fig. 3A) and a testing procedure (Fig. 3B) of 1,000 iterations. The (6 → 10 → 1) ANN model gave the best fit; the determination coefficient (r^2) was 0.83 for the training set and 0.80 for the testing set. The points in the scatterplots are well aligned along the diagonal of best prediction of coordinates 1 : 1.

Relationships between residuals and values predicted by the model show that the correlation coefficients were negligible and not significant ($r = 0.06$, $P = 0.316$, and $n = 300$ in the training set; $r = 0.14$, $P = 0.16$, and $n = 99$ in the testing set). We can thus consider residuals independent of the predicted values. The distribution of residuals was almost symmetrical and close to normal in the training set, with a mean value of 0.03 (SD = 0.16), and satisfied the assumption of normality in the testing set (mean = −0.03, SD = 0.07). To study the performance and the stability of ANN models, multiple runs were carried out with the training set of 300 observations and the testing set of 99 observations randomly chosen from the data set (Table 1). The results obtained from the five random training and testing sets showed that the determination coefficients were high for each of them (mean $r^2 = 0.81$, SD = 0.026 for training sets, and mean $r^2 = 0.77$, SD = 0.038 for testing sets). Despite the topographical heterogeneity and the large dimensions of the reservoir, results were satisfactory and showed a certain stability for the different random sampling points. The standard deviations of the correlation coefficients were very small in both the training and the testing sets.

The second test using the SM0 matrix is illustrated in Fig. 3C. Estimated values were sorted according to their stratum and showed two distinct patterns. In the surface stratum, estimated values were higher (mean = 0.60, SD = 0.13, $n = 156$) than in the underlying stratum (mean = 0.32, SD = 0.10, $n = 177$). The Mann–Whitney nonparametric statistical test showed a highly significant statistical difference ($U = 1{,}615$, $P < 0.001$) between FDs estimated in the two strata.

Multiple regression—The MR equations and determination coefficients were the following:

(1) Without transformation of the independent variables: in the training set (SM1train), $r^2 = 0.42$ (Fig. 4A), and in the testing set (SM1test), $r^2 = 0.55$ (Fig. 4B).

$$\log(\text{FD} + 1) = 3.197 - 4.5e^{-4}\text{DIS} - 0.102\text{OXY} - 2e^{-2}\text{DEP} - 8.1e^{-3}\text{SLO} - 9.4e^{-2}\text{STR} - 3.1e^{-2}\text{TEM}$$

(2) With nonlinear transformation of certain independent variables: in the training set (SM1train), $r^2 = 0.49$, and in the testing set (SM1test), $r^2 = 0.59$.

$$\begin{aligned}&\log(\mathrm{FD} + 1)\\&\quad = 3.542 - 4.3e^{-2}(\log(\mathrm{DIS})) - 0.818(\log(\mathrm{DEP}))\\&\qquad - 5.5e^{-2}(\log(\mathrm{SLO} + 1)) - 8.6e^{-2}\mathrm{OXY}\\&\qquad - 8.5e^{-2}\mathrm{STR} - 2.8e^{-2}\mathrm{TEM}\end{aligned}$$

For MR without transformation of the independent variables, all of the coefficients were significant ($P < 0.05$) except TEM. With transformation of some independent variables, only three of the six coefficients were significant ($P < 0.05$): DEP, STR, and OXY.

MR gave a rather good prediction of the data but was unable to learn (Fig. 4A) and predict (Fig. 4B) high values of FD, leading to its underestimation. Values of the determination coefficients both in training and testing sets indicate a clear improvement of the MR result after transformation of the variables. As this operation improves their linearity, we can conclude that nonlinear relationships exist between the dependent and independent variables. Thus, a GAM (i.e., nonparametric modeling method) was applied. The results showed a clear improvement of the correlation coefficients both in training and testing sets ($r^2 = 0.63$ in the training set [SM1train], and $r^2 = 0.70$ in the testing set [SM1test]). Nevertheless, high values of FD were systematically underestimated both in training (Fig. 4D) and testing sets (Fig. 4E). As a check, a method based on alternating least squares (ACE) was used to try to linearize the variables. After 60 iterations of the Transreg procedure, first with one knot and then with two, the optimal transformations finally derived yielded a determination coefficient $r^2 = 0.60$ with one knot and $r^2 = 0.67$ with two knots (Fig. 5A,B). Even though the determination coefficients were higher than for linear regression, they remained lower than those obtained using ANN and were very close to those obtained using GAM. MR after nonlinear transformation of the variables (i.e., GAM and ACE) always underestimated high values of FD (Figs. 4D,E, 5), and some aberrant values (i.e., negative values of FD) were predicted (Fig. 5B).

For comparing ANN, linear MR, and GAM, multiple runs were carried out on the same randomly chosen training and testing data sets from SM1 and with the same transformation of the dependent variable ($\log(x + 1)$ transformation of the FD). The results obtained from the five random training and testing sets showed that the determination coefficients were significant for all linear MR models (mean $r^2 = 0.42$, SD = 0.02 for training sets, and mean $r^2 = 0.51$, SD = 0.06 for testing sets), but the percentage of explained variability remained clearly lower for MR than for ANN and GAM results (Table 1). In the same way, between 31 and 49% of the variability is unexplained by the model using GAM (mean $r^2 = 0.66$, SD = 0.02 for training sets, and mean $r^2 = 0.59$, SD = 0.05 for testing sets), whereas ANN models always explained >70% of the variability in both training and testing procedures (mean $r^2 = 0.81$, SD = 0.03 for training sets, and mean $r^2 = 0.77$, SD = 0.04 for testing sets). Finally, the linear MR and GAM tests on the SM0 data set are represented in Fig. 4C,F. Estimated values were sorted according to their stratum and show two groups of estimated FD values corresponding to the upper (mean = 0.64, SD = 0.18, $n = 156$ for linear MR, and mean = 0.64, SD = 0.16, $n = 156$ for GAM) and the underlying strata (mean = 0.33, SD = 0.17, $n = 177$ for linear MR, and mean = 0.36, SD = 0.13, $n = 177$ for GAM). The Mann–Whitney nonparametric test shows a highly significant statistical difference ($U = 2{,}975$, $P < 0.001$ for linear MR, and $U = 1{,}871$, $P < 0.001$ for GAM) between the FD estimated in the two strata.

Sensitivity of the variables—Concerning ANN, the results of Garson's algorithm stress the relative contribution of the topographical variables in the model, with contributions of roughly 25% for DIS, 20% for DEP, and 15% for SLO. Moreover, the vertical fish distribution across the epilimnion (STR) contributes about 15%. Finally, among physical–chemical factors, only TEM makes an important contribution (ca. 15%), whereas OXY contributes <11%.

Applying Lek's algorithm, the influence of the six independent environmental variables on spatial fish distribution in the ANN model is illustrated by six curves (Fig. 6B).These show the influence of the six independent variables on the dependent variable in ANN modeling. Figure 6B shows four types of sensitivity (or contribution) curves: (1) exponentially decreasing contribution: DIS and DEP—the number of fish is maximal for low values of these two independent variables and then decreases rapidly down to very low FDs for 250 m from the bank and 15-m total depth; (2) linear decreasing contribution: STR—most fish are located in the upper stratum, i.e., 1–6 m; (3) linear increasing contribution: TEM—the FD is maximal for high values of the independent variable; and (4) weak contribution: SLO and OXY—the contribution of these variables hardly alters over their range. The profile can be thought of as a ''horizontal line.''

Using GAM, the lowess smoothing function and the observed FD (after $\log(x + 1)$ transformation) are reported in Fig. 6A. The response obtained for each predictor appears to be correctly related to the observed values. The trends of the lowess smoothing curves are the same as those obtained for ANN sensitivity analysis using Lek's algorithm. However, these trends are more clearly defined by using ANN sensitivity analysis than by using GAM for several variables such as DIS, DEP, and TEM.

To conclude, the four most important variables identified by ANN and GAM for the determination of the fish distribution are DIS, DEP, STR, and TEM.

Discussion

The spatial fish distribution studied here has been reliably fitted by ANN to the easily measured environmental characteristics of the points sampled in the lake. Thus, spatial variations of fish distribution in Lake Pareloup are strongly connected to a set of six environmental variables.

The main processes that determine fish distribution and diversity can be approximated by linear functions only to a limited extent, even using simple (e.g., logarithmic) trans-

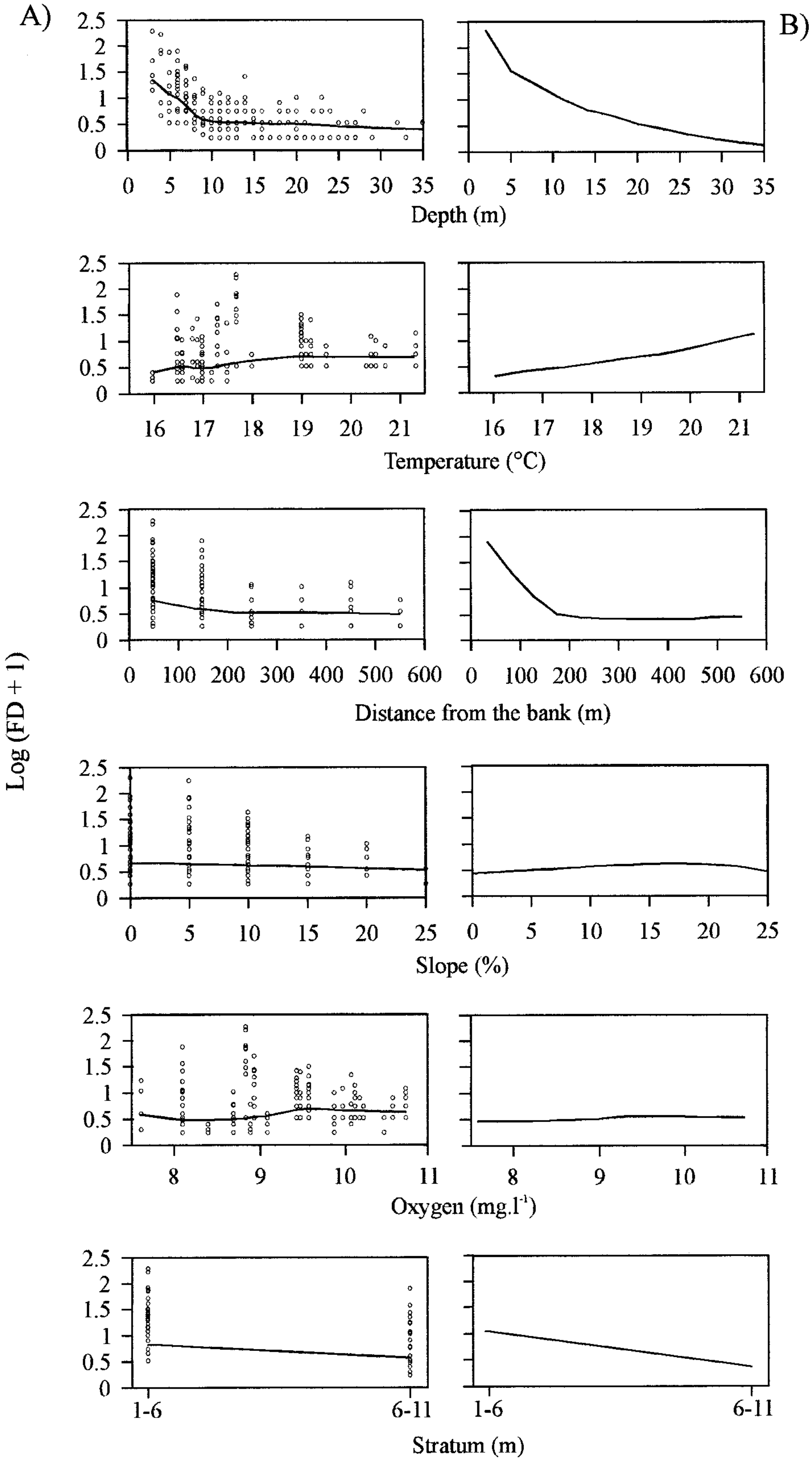

Fig. 6. Contribution profiles (sensitivity analysis) of each of the six independent variables for the prediction of FD using GAM (A) and ANN (B). The values cover the whole range of variation of each of the independent variables tested. For GAM (A), observed values of FD were plotted vs.

formations of variables to linearize their distribution. Although complex transformation of the variables using nonparametric modeling methods (ACE and GAM) clearly improves the determination coefficient, it still remains lower than that obtained by ANN. Moreover, such models are not able to reproduce the behavior of real systems when very low or high values of the variables are considered (Lek et al. 1996*b*). In ecology, models based on regression principles have been proposed by several authors (*see references in introductory material*). To improve results, nonlinear transformations of independent and/or dependent variables have frequently been used (Faush et al. 1988; Hakanson 1996). However, despite these transformations, the results often remained insufficient (low percentage of explained variance). In contrast, ANN with only one hidden layer can model nonlinear systems in ecology without transformation of the data (Goh 1995; Lek et al. 1996*b;* Scardi 1996). Of course, to model more complex systems, there is a need for complex networks (more units in the hidden layer or more than one hidden layer), adequate training, and larger data sets.

The use of backpropagation of the ANN weights led us to develop stochastic tools to predict the fish distribution from the environmental characteristics of a lake. The selection of input variables introduced into the modeling procedures, their ecological significance, and the constitution of testing sets of data to assess the performance of the model are important elements for this type of approach (Fausch et al. 1988). In the present work, the ANN backpropagation procedure gave very high correlation coefficients between observed and predicted values, both in training and testing sets, which were always clearly significant.

The tests performed on submatrix SM0 stress the enhanced ability of ANN compared with MR to discriminate between the two strata. Even though parametric and nonparametric MR can reveal a significant statistical difference between FD in the two strata, Fig. 4C,F shows a large range of responses inside the same stratum and predicts certain aberrant (negative) values of FD (Fig. 4C). Moreover, the test performed using ANN provided interesting insights into the potential fish distribution in the studied area (Fig. 3C). We can hypothesize on the basis of the environmental variables taken into account that, among the units where no fish were detected, some constitute a potentially better habitat for fish than others.

Figure 3C presents two distinct FD patterns for the STR variable and reveals that among the units without fish, those located in the surface stratum constitute a potentially better habitat for fish than those located in the underlying stratum. Some hypotheses can be formulated to explain the absence of fish from these units: (1) shoaling could have caused a patchy distribution inside homogeneous physical–chemical areas; and (2) the FDs in the surface stratum could have been underestimated because of fish escaping from the motor-driven boat (Schultz 1988). This could explain the nil FDs recorded in some areas that appeared to offer a favorable habitat for fish and could also explain the difference between expected and observed values in the ANN approach. Moreover, the SM0 test validates the efficiency of ANN. Predicted values are not real values, but, on the basis of the training procedure, the ANN has defined the potential of each unit of the SM0 set to receive a certain FD. This proves that ANN models are able to reproduce ecological factors, whereas MR shows many shortcomings, even if the determination coefficient remains significant.

A theoretical disadvantage of ANN models is that their parameters do not provide information about the relative importance of the independent variables (although this is not true when composite variables are used). However, this problem can be solved by performing sensitivity analysis of the ANN. Garson (1991) and Goh (1995) have proposed methods for interpreting neural network connection weights to illustrate the importance of explanatory variables inside the ANN. These studies demonstrate the potential of ANN approaches to explain nonlinear interactions between variables in complex systems and propose a procedure for partitioning the connection weights to determine the relative importance of the various input variables. In ecology, Lek et al. (1996*a,b*) proposed an algorithm allowing the visualization of the profiles of explanatory variables. In addition to the predictive value of the model, we attempted to detect, by a simple simulation method, the sensitivity of the different variables.

The ANN models clearly show how each of the variables acts in a nonlinear way on the general lake ecosystem. Comparison between GAM and ANN sensitivity showed the same trends as those visualized by plotting each ecological variable vs. the observed response (FD). Moreover, these trends are more clearly underlined by ANN sensitivity curves than those obtained by the lowess smoothing function used in the GAM (Fig. 6). However, both GAM and ANN sensitivity analysis underlined the influence of several environmental variables on the heterogeneity of fish distribution in the lake. Fish were mainly located in the surface stratum and in the warm shallow littoral areas, whereas the underlying waters and the deep, cold, and distant areas were systematically avoided. This information, provided by the sensitivity analysis of the variables, is supported by ecological factors. The heterogeneous distribution observed in Lake Pareloup is supported by the ecological trends reported in the literature (Rossier 1995; Fischer and Eckmann 1997) and is closely linked to four of the six environmental variables taken into account. Moreover, most whole-lake studies describe the fish distribution from a global point of view, considering only two kinds of macrohabitats, the pelagic and the littoral zones, without more precise localization (Bohl 1980; Goldspink 1990; Eklöv 1997). Other works focus only on either pelagic (Hamrin 1986; Jurvelius et al. 1988) or littoral (Rossier 1995; Fischer and Eckmann 1997) areas. In the same way, fish distribution is commonly defined using only one or two chemical or topographical variables such as

←

each of the independent variables. On the same plots, the lowess smoothing function with $f = 0.50$ is represented by a solid line. No lowess function is plotted for the GAM stratum profile because there are only two strata; this variable was fitted by a linear function.

water temperature (Hamrin 1986; Imbrock et al. 1996; Kubecka and Wittingerova 1998), dissolved oxygen concentration and water conductivity (Matthews et al. 1985), and depth and distance from the bank (Rossier 1995; Fischer and Eckmann 1997). Even if these studies based on hydroacoustics and net catches provide insights into the local habitats of fish, they are unable to extend the results to the scale of a whole lake and fail to define general trends in fish distribution, for our results show that fish distribution is ruled by a complex combination of several variables acting mainly in a nonlinear way.

From an ecological point of view, the correct choice of habitat is of crucial importance for individual survival and therefore determines local species distribution (Rosenzweig 1991). Fish distribution permanently seeks a trade-off between available habitat and the necessity to accomplish vital functions (Lévéque 1995), leading to habitat partitioning (Schoener 1974). Therefore, the heterogeneity of the fish distribution assessed using ANN probably results from these trade-offs.

The fish community recorded in Lake Pareloup was composed of 15 species, but it was mainly represented by roach (*Rutilus rutilus* L.), representing numerically >84% of the adult fish community (Richeux et al. 1994). Roach feed preferentially during the day (Persson 1983; Jamet et al. 1990) and generally present an opportunistic feeding behavior: they feed on plankton, macrophytes, benthos, and sediment (Niederholzer and Hofer 1980; Persson 1983, 1987; Jamet et al. 1990; Michel and Oberdorff 1995). Therefore, roach are located, on the one hand, in the upper water stratum (1–6 m), where plankton density is the greatest in Lake Pareloup (Francisco and Rey 1994), and, on the other hand, near the bottom, in warm shallow areas, mainly composed of mud with patches of filamentous algae and macrophytes, which constitute an important food item for roach during summer (Prejs and Jackowska 1978; Persson 1983; Michel and Oberdorff 1995). Finally, the predation pressure by piscivorous fish (pike-perch, large perch, and pike), which numerically represent >9% of the adult fish community (Richeux et al. 1994), could restrict roach distribution in the lake (Eklöv 1997). We denote a clear avoidance of deep areas (stratum = 6–11 m), which may be considered predation avoidance of pike-perch (Brabrand and Faafeng 1994). Consequently, most of the essentially roach fish community exhibits a distinctly patchy distribution in the surface waters of the littoral zone. Finally, this shows that overall trends in fish distribution can be easily assessed using a few pertinent environmental variables and provides insights into the ecological meaning of fish distribution on the scale of a whole lake.

The influence of environmental variables on fish distribution, assessed using sensitivity analysis of the variables supplied by ANN models, is in accordance with ecological factors reported in previous studies (*see above references*). Thus, ANN models are able to reproduce the operation of real systems on the basis of the ecological variables introduced in the model. Moreover, the predictive power of ANN overpasses the capabilities of more common techniques, even though GAM gave acceptable results, without reaching the same performance as ANN from a predictive and, to a lesser extent, an explanatory point of view. Consequently, ANN models can be used either as a predictive tool or as an explanatory tool where parametric and nonparametric regression methods are quite limited.

To conclude, the ANN modeling approach used here is a fast and flexible way to incorporate multiple input parameters into a single model. It is this ability to deal with multiple information sources that provides the power of this approach, which results in a significant improvement in ANN modeling over conventional techniques.

References

BAKER, J. R., AND L. J. PAULSON. 1983. The effects of limited food availability on the stripped bass fishery in Lake Mead, p. 551–561. *In* V. D. Adams and V. A. Lamarra [eds.], Aquatic resources management in Colorado River ecosystem. Ann Arbor Science Publications.

BARAN, P., S. LEK, M. DELACOSTE, AND A. BELAUD. 1996. Stochastic models that predict trout population densities or biomass on microhabitat scale. Hydrobiologia **337:** 1–9.

BOHL, E. 1980. Diel pattern of pelagic distribution and feeding in planktivorous fish. Oecologia **44:** 368–375.

BRABRAND, A., AND B. FAAFENG. 1994. Habitat shift in roach (*Rutilus rutilus*) induced by the introduction of pike perch (*Stizostedion lucioperca*). Verh. Int. Verein. Limnol. **25:** 2123.

BREIMAN, L., AND J. H. FREIDMAN. 1985. Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc. **77:** 580–619.

BURCZYNSKI, J. J., P. H. MICHALETZ, AND G. M. MARRONE. 1987. Hydroacoustic assessment of the abundance and distribution of rainbow smelt in Lake Oahe. N. Am. J. Fish. Manag. **7:** 106–116.

CASSELMAN, F. L., D. F. FREEMAN, D. A. KERRIGAN, S. C. LANE, D. M. MAGLEY, N. H. MILLSTORM, AND C. R. ROY. 1994. A neural network-based underwater acoustic application, p. 3409–3414. *In* Proceedings of the International Conference on Neural Networks, IEEE.

CLEVELAND, W. S. 1979. Robust locally-weighted regression and scatterplot smoothing. J. Am. Stat. Assoc. **74:** 829–836.

EGGERS, D. M. 1978. Limnetic juvenile feeding behavior of juvenile sockeye salmon in Lake Washington and predator avoidance. Limnol. Oceanogr. **23:** 1114–1125.

EKLÖV, P. 1997. Effects of habitat complexity and prey abundance on the spatial and temporal distributions of perch (*Perca fluviatilis*) and pike (*Esox lucius*). Can. J. Fish. Aquat. Sci. **54:** 1520–1531.

FAUSH, K. D., C. L. HAWKES, AND M. G. PARSONS. 1988. Models that predict the standing crop of stream fish from habitat variables: 1950–85. General Technical Rep. PNW-GTR. U.S. Department of Agriculture.

FISCHER, P., AND R. ECKMANN. 1997. Spatial distribution of littoral fish species in a large European lake, Lake Constance, Germany. Arch. Hydrobiol. **140:** 91–116.

FOOTE, K. G. 1982. Optimizing copper spheres for precision calibration of hydroacoustic equipment. J. Acoust. Soc. Am. **71:** 612–616.

FRANCISCO, P., AND J. REY. 1994. Etude du peuplement zooplanctonique de la retenue de Pareloup (Aveyron, France). Hydroecol. Appl. **6:** 175–196.

GARSON, G. D. 1991. Interpreting neural network connection weights. Artif. Intel. Expert. **6:** 47–51.

GEMAN, S., E. BIENENSTOCK, AND R. DOURSAT. 1992. Neural networks and the bias/variance dilemma. Neural Comput. **4:** 1–58.

GOH, A. T. C. 1995. Back-propagation neural networks for modeling complex systems. Artif. Intel. Eng. **9:** 143–151.

GOLDSPINK, C. R. 1990. The distribution and abundance of young (I+–II+) perch, *Perca fluviatilis* L., in a deep eutrophic lake, England. J. Fish Biol. **36:** 439–447.

GUÉGAN, J. F., S. LEK, AND T. OBERDORFF. 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature **391:** 382–384.

HAKANSON, L. 1996. Predicting important lake habitat variables from maps using modern modeling tools. Can. J. Fish. Aquat. Sci. **53:** 364–382.

HAMRIN, S. F. 1986. Vertical distribution and habitat partitioning between different size classes of Vendace (*Coregonus albula*), in thermally stratified lakes. Can. J. Fish. Aquat. Sci. **43:** 1617–1625.

HASLER, A. D., AND J. R. VILLEMONTE. 1953. Observations on the daily movements of fishes. Science **118:** 321–322.

HASTIE, T. J., AND R. J. TIBSHIRANI. 1990. Generalized additive models. Chapman and Hall.

HRUSKA, V. 1989. Abundance and the spatial distribution of fish echotraces in the Rimov reservoir. Arch. Hydrobiol. Beih. Ergebn. Limnol. **33:** 615–617.

IMBROCK, F., A. APPENZELLER, AND R. ECKMANN. 1996. Diel and seasonal distribution of perch in Lake Constance: A hydroacoustic study and *in situ* observations. J. Fish Biol. **49:** 1–13.

JAGER, J. C., AND C. W. N. LOOMAN. 1995. Data collection, p. 10–28. *In* R. G. H. Jongman, C. J. F. ter Braak, and O. F. R. Van Tongeren [eds.], Data analysis in community and landscape ecology. Cambridge Univ. Press.

JAMES, F. C., AND C. E. MCCULLOCH. 1990. Multivariate analysis in ecology and systematics: Panacea or Pandora's box? Annu. Rev. Ecol. Syst. **21:** 129–166.

JAMET, J. L., P. GRES, N. LAIR, AND G. LASSERRE. 1990. Diel feeding cycle of roach (*Rutilus rutilus* L.) in eutrophic Lake Aydat (Massif Central, France). Arch. Hydrobiol. **118:** 371–382.

JURVELIUS, J., T. LINDEM, AND T. HEIKKINEN. 1988. The size of vendace, *Coregonus albula* L., stock in a deep lake basin monitored by hydroacoustic methods. J. Fish Biol. **32:** 679–687.

KOHAVI, R. 1995. A study of the cross-validation and bootstrap for accuracy estimation and model selection, p. 1137–1143. *In* Proceedings of the International Joint Conference on Artificial Intelligence.

KUBECKA, J., AND M. WITTINGEROVA. 1998. Horizontal beaming as a crucial component of acoustic fish stock assessment in freshwater reservoirs. Fish. Res. **35:** 99–106.

LEK, S., A. BELAUD, P. BARAN, I. DIMOPOULOS, AND M. DELACOSTE. 1996*a*. Role of some environmental variables in trout abundance models using neural networks. Aquat. Living Resour. **9:** 23–29.

———, M. DELACOSTE, P. BARAN, I. DIMOPOULOS J. LAUGA, AND S. AULAGNER. 1996*b*. Application of neural networks to modeling nonlinear relationships in ecology. Ecol. Model. **90:** 39–52.

———, I. DIMOPOULOS, AND A. FABRE. 1996*c*. Predicting phosphorus concentration and phosphorus load from watershed characteristics using backpropagation neural networks. Acta Oecol. **17:** 43–53.

LÉVÊQUE, C. 1995. L'habitat: Être au bon endroit au bon moment? Bull. Fr. Pêche Piscic. **337/338/339:** 9–20.

MACLENNAN, D. N., and E. J. Simmonds. 1992. Fisheries acoustics. Chapman and Hall.

MASTRORILLO, S., S. LEK, F. DAUBA, AND A. BELAUD. 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biol. **38:** 237–246.

MATTHEWS, W. J., L. G. HILL, AND S. M. SCHELLHAASS. 1985. Depth distribution of Striped Bass and other fish in Lake Texoma (Oklahoma–Texas) during summer stratification. Trans. Am. Fish. Soc. **114:** 84–91.

MICHEL, P., AND T. OBERDORFF. 1995. Feeding habits of fourteen European freshwater fish species. Cybium **19:** 5–46.

NIEDERHOLZER, R., AND R. HOFER. 1980. The feeding of roach (*Rutilus rutilus*) and rudd (*Scardinius erythrophtalmus*). 1. Studies in natural populations. Ekol. Pol. **25:** 241–255.

NORUSIS, M. J. 1993. SPSS for Windows. Base system user's guide, release 6.0. SPSS.

O'BRIEN, W. J., B. LOVELESS, AND D. WRIGHT. 1984. Feeding ecology of young white crappie in a Kansas reservoir. N. Am. J. Fish. Manag. **4:** 341–349.

PERSSON, L. 1983. Food consumption and the significance of detritus/algae to the intraspecific competition in roach (*Rutilus rutilus*) in a shallow eutrophic lake. Oikos **41:** 118–125.

———. 1987. Effects of habitat and season on competitive interactions between roach (*Rutilus rutilus*) and perch (*Perca fluviatilis*). Oecologia **73:** 170–177.

POFF, N., S. TOKAR, AND P. JOHNSON. 1996. Stream hydrological and ecological responses to climate change assessed with an artificial neural network. Limnol. Oceanogr. **41:** 857–863.

PREJS, A., AND H. JACKOWSKA. 1978. Lake macrophytes as the food of roach (*Rutilus rutilus* L.) and rudd (*Scardinius erythrophtalmus* L.). I. Species composition and dominance relations in the lake and food. Ekol. Pol. **26:** 429–438.

RICHEUX, C., J-F. NOGUES, J-N TOURENQ, AND B. ARAGON. 1994. Inventaire piscicole de la retenue hydroéléctrique de Pareloup (Aveyron, France) lors de la vidange de Juin 1993. Essai d'un nouveau systéme d'acquisition et de traitement des signaux d'un échosondeur. Hydroécol. Appl. **6:** 197–226.

ROSENZWEIG, M. L. 1991. Habitat selection and population interactions: The search for mechanisms. Am. Nat. **137:** 5–28.

ROSSIER, O. 1995. Spatial and temporal separation of littoral zone fishes of Lake Geneva (Switzerland–France). Hydrobiologia **300/301:** 321–327.

RUMELHART, D. E., G. E. HINTON, AND R. J. WILLIAMS. 1986. Learning representations by back-propagating error. Nature **323:** 533–536.

[SAS] STATISTICAL ANALYSIS SYSTEMS. 1988. SAS technical report P-179. Additional SAS/STAT procedures, release 6.03. Statistical Analysis Systems Institute.

SCARDI, M. 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Mar. Ecol. Prog. Ser. **139:** 289–299.

SCHOENER, T. 1974. Resource partitioning in ecological communities. Science **185:** 27–39.

SCHULTZ, H. 1988. An acoustic fish stock assessment in the Bautzen Reservoir. Limnologica (Berlin) **19:** 61–70.

SMITH, M. 1993. Neural networks for statistical modeling. Van Nostrand Reinhold.

TER BRAAK, C. J. F., AND C. W. N. LOOMAN. 1995. Regression, p. 29–77. *In* R. G. H. Jongman, C. J. F. ter Braak, and O. F. R. Van Tongeren [eds.], Data analysis in community and landscape ecology. Cambridge Univ. Press.

THORNE, R. E. 1983. Hydroacoustics, p. 239–259. *In* L. A. Nielsen and D. L. Johnson [eds.], Fisheries techniques. American Fisheries Society.

YOUNG, F. W. 1981. Quantitative analysis of qualitative data. Psychometrika **46:** 357–388.

Received: 10 June 1998
Accepted: 18 February 1999
Amended: 25 March 1999