Predicting local fish species richness in the Garonne River basin

Modèle prédictif de la richesse spécifique locale des poissons du bassin de la Garonne

Sylvain Mastrorillo[a]*, Francis Dauba[b], Thierry Oberdorff[c], Jean-Francois Guégan[d], Sovan Lek[a]

[a] CNRS UMR 5576, Cesac, université Paul-Sabatier, bât. IVR3, 118, route de Narbonne, 31062 Toulouse cedex, France
[b] Laboratoire ingénierie agronomique, Équipe environnement aquatique et aquaculture, INP-Ensat, avenue de l'Agrobiopole, BP 107, Auzeville Tolosane, 31326 Castanet-Tolosan, France
[c] Laboratoire d'ichtyologie générale et appliquée, Muséum national d'histoire naturelle, 43, rue Cuvier, 75231 Paris cedex 05, France
[d] Orstom, CNRS UMR 5556, Station méditerranéenne de l'environnement littoral, université Montpellier-II, 1, quai de la Daurade, 34200 Sète, France

(Received 9 December 1997, accepted after revision 22 December 1997)

Abstract – The aim of this work was to predict local fish species richness in the Garonne river basin using three environmental variables (distance from the source, elevation and catchment area). Commonly, patterns of fish species richness have been investigated using simple or multi-linear statistical models. Here, we used backpropagation of artificial neural networks (ANNs) to develop stochastic models of local fish diversity. Two independent data collections were used, the first one to build and test the model; the second one to validate the model. Correlation coefficients between observed values and predicted values both in the testing and the validation procedures were highly significant ($r = 0.904$, $P < 0.001$ and $r = 0.822$, $P < 0.001$, respectively). The ANN model obtained using only three environmental variables succeeded in explaining ca 70 % of the total variation in local fish species richness. Through these findings, ANNs can be seen as a powerful predictive tool compared to traditional modelling approaches. (© Académie des sciences / Elsevier, Paris.)

local species richness / fish / Garonne river basin / environmental variables / artificial neural networks

Résumé – Ce travail a consisté à prédire la richesse spécifique locale de poissons dans le bassin de la Garonne à partir de trois variables environnementales : la distance à la source, l'altitude et la surface du bassin versant. Le plus souvent, la richesse spécifique d'un bassin est étudiée en utilisant des modèles de régression linéaire simple ou multiple. Dans notre travail, un réseau de neurones a été utilisé pour développer un modèle prédictif de la richesse spécifique. Deux bases de données indépendantes ont été utilisées, une pour construire et tester le modèle, l'autre pour le valider. Les coefficients de corrélation obtenus pour le test et la validation sont élevés et hautement significatifs (respectivement, $r = 0{,}904$, $p < 0{,}001$ et $r = 0{,}822$, $p < 0{,}001$). Près de 70 % de la variation de la richesse est expliquée par les trois variables environnementales. À travers ces résultats, les réseaux de neurones artificiels peuvent être considérés comme un puissant outil prédictif face aux approches de modélisations plus traditionnelles. (© Académie des sciences / Elsevier, Paris.)

richesse spécifique locale / poissons / bassin de la Garonne / variables environnementales / réseaux de neurones artificiels

Note communicated by Henri Décamps.

*Correspondence and reprints
E-mail: mastrori@cict.fr

Version abrégée

Contrairement aux autres grands fleuves français (Rhône, Seine, Loire), peu d'études ont été réalisées pour décrire les communautés de poissons à l'échelle du bassin de la Garonne. Il nous a semblé d'autant plus intéressant de développer un telle recherche que le bassin de la Garonne est moins perturbé que les autres par des pollutions d'origine industrielle. Cependant, l'édification de nombreux barrages sur la Garonne et ses affluents a modifié au cours du temps les régimes hydrologiques et favorisé le phénomène de fragmentation des communautés animales et végétales dans le fleuve principale mais aussi dans toute la plaine alluviale.

De nombreux travaux sur la répartition spatiale des poissons à l'échelle d'un bassin versant montrent que la richesse spécifique locale (la richesse spécifique sur un site) change de façon prévisible le long du continuum fluvial. Il est généralement admis que la richesse spécifique croît avec une augmentation en taille du cours d'eau pour atteindre un maximum (plateau) à partir des zones intermédiaires (pour éventuellement décroître dans les zones les plus aval).

Ces études sur l'évolution de la richesse spécifique ont été généralement réalisées en utilisant les méthodes de régression linéaire simple ou multiple. Dans ce travail, nous proposons un modèle prédictif de la richesse spécifique locale dans le bassin de la Garonne à partir de trois variables environnementales simples (distance à la source, altitude et surface du bassin versant). Pour ce faire, nous avons utilisé les réseaux de neurones artificiels dont la capacité de prédiction a été démontrée dans divers domaines de l'écologie.

Pour construire et tester le modèle, nous avons utilisé des données de pêches électriques standardisées, réalisées en période d'étiage, sur 207 stations réparties sur l'ensemble du bassin de la Garonne. Pour chaque station, la distance à la source, l'altitude et la surface du bassin versant sont relevées. Ce jeu de données a été scindé en un ensemble d'apprentissage de 155 observations (75 % des stations) pour construire le modèle et un ensemble de test de 52 observations (25 % des stations) pour tester le modèle. Les réseaux de neurones utilisés sont de type « *feed-forward* » à trois couches avec une fonction de transfert de type sigmoïde. Les coefficients de corrélation entre les valeurs observées et les valeurs estimées (apprentissage) ou prédites (test) sont respectivement de 0,935 et 0,904, et sont hautement significatifs ($p < 0,001$). L'étude des relations entre résidus et valeurs estimées ou prédites montre une indépendance entre ces valeurs.

La validation de ces résultats a été entreprise sur un deuxième jeu de données (72 stations de pêche électrique), indépendant tant du point de vue de son origine (pêches réalisées par un organisme différent) que des stations prospectées, dans le but de déterminer la qualité prédictive du modèle. À notre connaissance, c'est la première fois en écologie que la validation d'un modèle neural est réalisée sur un jeu de données indépendant. Le coefficient de corrélation entre les valeurs observées et les valeurs prédites est élevé ($r = 0,822$) et hautement significatif ($p < 0,001$). Exception faite de quelques valeurs surestimées ou sous-estimées, la majorité des points sont parfaitement prédits par le modèle.

Les études existantes sur la richesse spécifique de poissons réalisées dans d'autres grands fleuves expliquent environ 50 % de la variabilité totale de cette richesse (et cela quel que soit le nombre de variables prédictives prises en compte) tandis que le modèle de réseau de neurones utilisé dans notre étude explique plus de 70 % avec seulement trois variables.

Une approche expérimentale a été employée pour déterminer la réponse du modèle pour chaque variable environnementale. L'influence de ces variables sur la richesse spécifique a été illustrée sous forme de courbes (ou profils) de sensibilité. On montre ainsi que la richesse spécifique présente une relation logarithmique croissante avec la distance à la source, une relation sigmoïdale décroissante avec l'altitude et une relation sigmoïdale croissante avec la surface du bassin versant.

Nos résultats confirment l'évolution classique de la richesse spécifique le long du gradient amont–aval d'un bassin hydrographique. Le nombre d'espèces augmente régulièrement avec la taille du cours d'eau pour atteindre un plateau à partir des zones intermédiaires. À travers cet exemple, les réseaux de neurones artificiels peuvent être considérés comme une méthode prédictive puissante face aux méthodes traditionnelles de modélisation. Ils pourraient être utilisés pour comprendre et expliquer les phénomènes non linéaires agissant dans la structuration des communautés animales et végétales.

1. Introduction

Many published studies on fish assemblage structure in rivers suggest that local species richness (species richness at a site) shifts in a predictable fashion along an upstream–downstream gradient. Large scale basin investigations suggest that fish species richness generally increases with river size (measured as a function of gradient [1, 2], river width [3], stream order [4, 5], distance from sources [3] and catchment area [6, 7]), and then reaches an asymptote (or sometimes decreases) in downstream areas [6, 8–11]. The reasons for these findings are often attributed to a downstream increase in habitat diversity [12–14] and/or to the degree of stability of the physical environment (e.g. discharge stability) which usually tends to rise with increasing river size [4].

Here, we analysed patterns of fish species richness in the Garonne river basin. Relatively little work has been carried out to describe fish assemblages in this river compared to other large rivers of France, e.g. the Seine river [6, 7, 10] and the Rhône river [6, 15], except that by Lim et al.

[16] who analysed variation of fish assemblage structure among six sites located in the middle course of the river.

Commonly, patterns of fish species richness have been investigated using simple or multi-linear statistical models [6, 13, 15, 17]. However, recent works have demonstrated that the use of artificial neural networks (ANNs) hold great promise for these kinds of data sets [18–21]. ANNs are known for their capacity to process non-linear relationships between variables which usually confound classical statistical methods for species richness prediction. ANNs can both improve performance by training and define complex relationships between independent and dependent variables that are not apparent using other methods, e.g. [22, 23].

In this paper, we 1) examined the capacity of neural network models to predict local fish species richness (LSR) in the Garonne River basin using three environmental variables (distance from source, elevation and catchment area), 2) validated the model on an independent data set, 3) identified the importance of these predictive variables in the model, 4) discussed the potential applications of ANNs method in conservation ecology.

2. Materials and methods

2.1. Study area

The River Garonne has its source in the Maladetta Glacier (Spain), and it slopes from the southeast to the northwest, where it reaches the Atlantic ocean through the Gironde estuary. The River Garonne drains an area of about 57 000 km² and its total length is 525 km. Mean annual discharge amounts to about 545 $m^3.s^{-1}$. Comparing with other French rivers (e.g. the Seine river and the Rhône river), the Garonne river is less disturbed by industrial pollution. However, its natural flow has been modified by the presence of several dams, promoting in that way, animal and vegetal community fragmentation within the river channel and the alluvial floodplain [24, 25].

From south to northwest, topography and climate determine three great landscape types: the Pyrenean mountains with a pronounced relief, a vast, green hill zone of Piedmont, and the valley of the Garonne river with flooding zones and alluvial terraces. The oceanic influence predominates on the totality of the basin, but lessens to the southeast where it undergoes the Mediterranean influence with dry winds and weaker pluviometry.

2.2. Data collection

To build and to test the model, we used data from 207 fish sampling sites fairly evenly distributed across the Garonne river basin (*figure 1*). Data for LSR (including successfully introduced species) were collected between 1986 and 1996 by the Laboratoire d'ingénierie agronomique, Ensat. All sites were sampled once by electrofishing, during low-flow periods, and using standardized methods.

To validate the model, we used an independent data set constituted by 72 different sites sampled by the Conseil supérieur de la pêche from 1985 to 1995 and using the same sampling procedures.

At each studied site, three environmental variables were recorded: distance from source (DFS), elevation (ELE) and catchment area (CAA). The CAA at each site was measured with a digital planimeter on a 1/500 000 scale map of the Garonne river basin. The two other variables (DFS and ELE) were recorded on several 1/25 000 scale maps.

2.3. Methods of modelling

A three-layer backpropagation of ANNs (*figure 2*) was used with an input layer of three neurons (one for each of the environmental variables), a hidden layer of three neurons which represent the best compromise of bias and variance [26], and an output layer of one neuron to predict the LSR. The learning rate and momentum initially fixed at 0.01 and 0.95 are modified during the iterations according to the importance of the error between observed and estimated values. The initial weights for the linkages between neurons were randomly chosen within a range between –0.3 and 0.3. The backpropagation algorithm [26] was used to change these connection weights during the training procedure. Details of this algorithm can be found in refs [26]and [27]. The ANN analysis was realized on a compatible PC with Pentium® processor using a

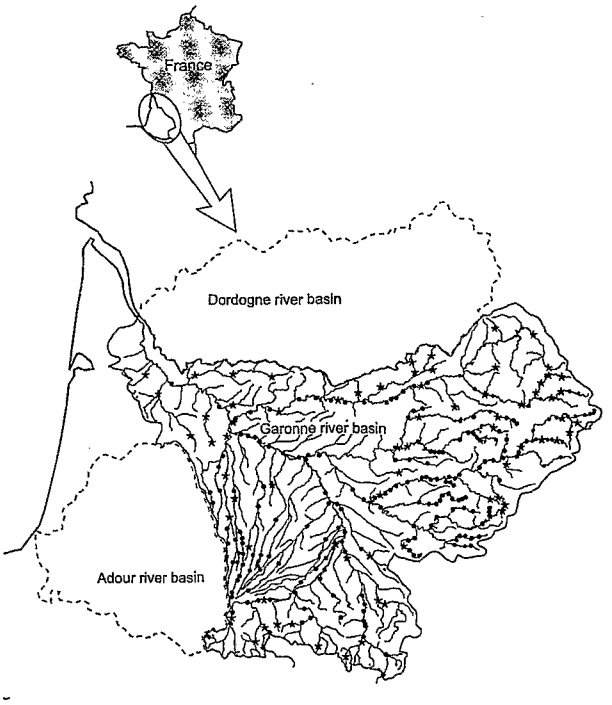

Figure 1. Distribution map of the collecting sites located on the Garonne river basin.

Black points represent the 207 fish sampling sites by the Laboratoire d'ingénierie agronomique, Ensat; black stars the 72 fish sampling sites by the Conseil supérieur de la pêche.

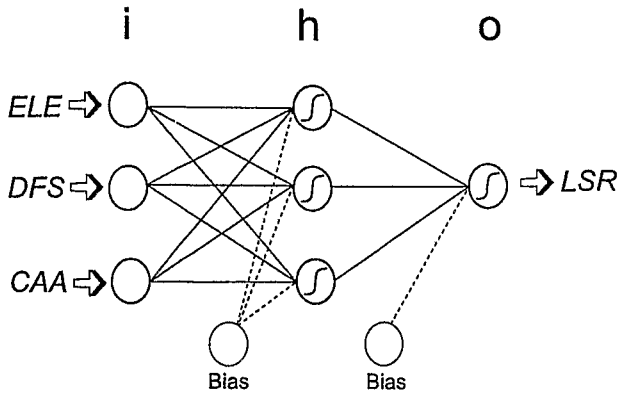

Figure 2. Representation of the typical three-layered feed-forward artificial neural network used.

Three input nodes (layer i) corresponding to the three independent variables (CAA: catchment area; DFS: distance from the source; ELE: elevation), three nodes on one hidden layer (layer h) and one output node (layer o) corresponding to the estimate of local fish species richness (LSR).

matricial software®, MATLAB with the Neural Network toolbox.

Modelling was carried out in three steps: 1) the first data matrix [207 rows (sites) × 4 columns (DFS, ELE, CAA and LSR)] were randomly divided into two sets of data and the first set, i.e. 75 % of the data or 155 records, was used to determine connection weights after 500 iterations of the training procedure; 2) the second set, i.e. 25 % or 52 records, was used to test the previous model; 3) the validation of our results was performed on the second independent data matrix [72 rows (sites) × 4 columns (DFS, ELE, CAA and LSR)] in order to determine the predictive quality of the model. An experimental approach was used to determine the response of the model to each of the input variables [18, 23].

3. Results

3.1. Training of the network during step 1

ANNs were trained by the backpropagation algorithm following the step one procedure (see above). The number of iterations was limited to 500 (best compromise between bias and variance) which is quite low in neural network modelling. This neural network configuration gives only 16 parameters (three input nodes × three hidden nodes + three hidden nodes × one output nodes + four biases).

The resulting correlation coefficient is 0.935 ($P < 0.001$) for the regression between observed and estimated values (*figure 3A*), the points are well aligned on the diagonal of the perfect fit line (co-ordinate 1:1). The relationship between residuals and estimated values

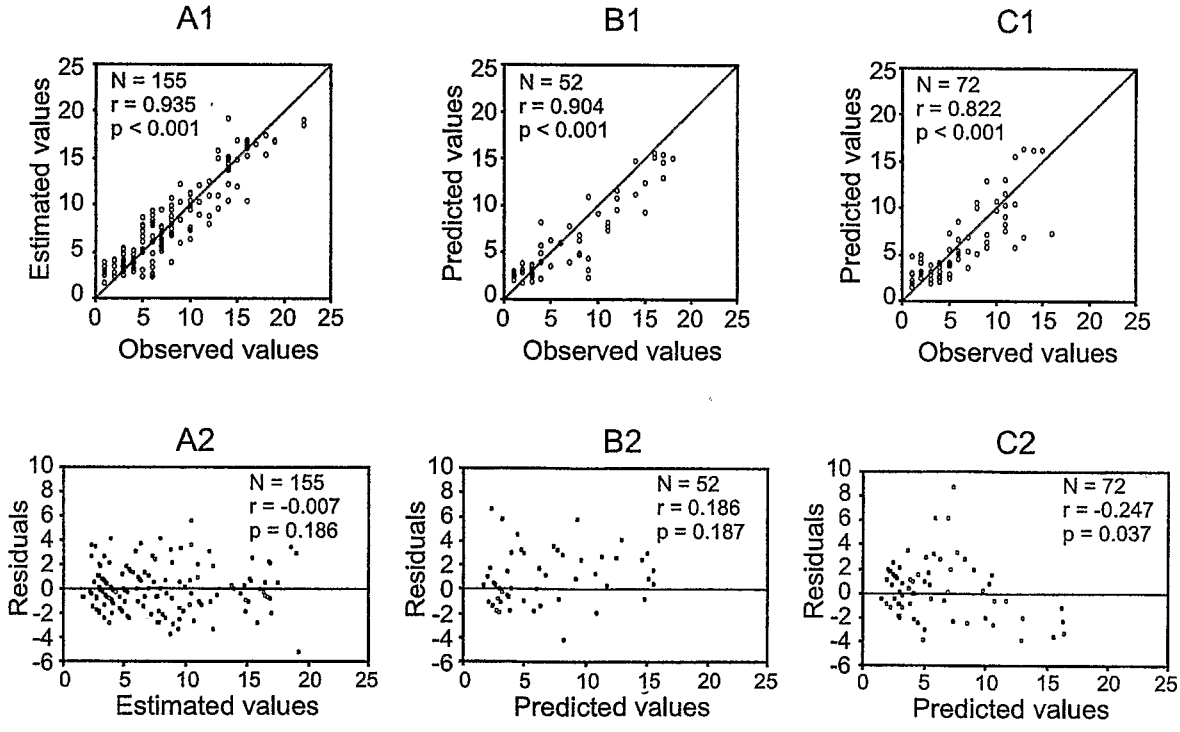

Figure 3. Results of the modelling.

A: Training (step 1). **B:** Testing (step 2). **C:** Validation (step 3).

1: Scatter plot of observed values versus estimated or predicted values: the solid line indicates the perfect fit line (for which y = x).

2: Relationship between residuals and estimated or predicted values. In step 3, the most under-estimated value correspond to a site located downstream an artificial lake with a concentration of lots of fish species.

shows complete independence of values (r = – 0.007, $P = 0.935$).

3.2. Testing of the network during step 2

The ANN procedure tested on the remaining 25 % (52 records) shows that the majority of records are aligned on the diagonal of co-ordinate 1:1 (*figure 3B*). The resulting correlation coefficient is 0.904 ($P < 0.001$). Relationship between residuals and estimated values shows complete independence (r = 0.186, $P = 0.187$).

3.3. Validation of the network during step 3

Interestingly, ANNs were validated on the totality of the second data matrix (72 records). Our results demonstrate the high predictive power of the model. In fact, the correlation coefficient between observed and predicted values is 0.822 ($P < 0.01$). Over/underestimates of some values (especially for sites where LSR > 10) may be observed (*figure 3C*) which results in some residuals showing some dependence (r = –0.247, $P = 0.04$).

3.4. Importance of the three environmental variables and neural network sensitivity

By applying Garson's algorithm [28], distance from the source is the major contributing variable to the model, but contribution percentages of each environmental variable are very similar (i.e. 39 % for DFS, 33 % for ELE and 28 % for CAA).

The influence (or sensitivity profile) of the independent variables on the fish species richness (LSR) in the ANN modelling determined by Lek's algorithm [18, 23] is illustrated in *figure 4*.

– There is a sigmoïd decrease between LSR and ELE (*figure 4* – top). After constant values of LSR up to 250 m, LSR decreases rapidly as the value of elevation increases up to 500–1 000 m. Then, LSR decreases very slowly to reach less than two species up to 2 000 m.

– There is a logarithmic increase between LSR and DFS (*figure 4* – middle). First, LSR rises rapidly with DFS to reach an asymptote for DFS > 200 km.

– There is a sigmoïd increase between LSR and CAA (*figure 4* – bottom). LSR increases slowly with CAA up to 10 000 km², and then LSR increases rapidly with CAA to reach an asymptote for CAA > 30 000 km².

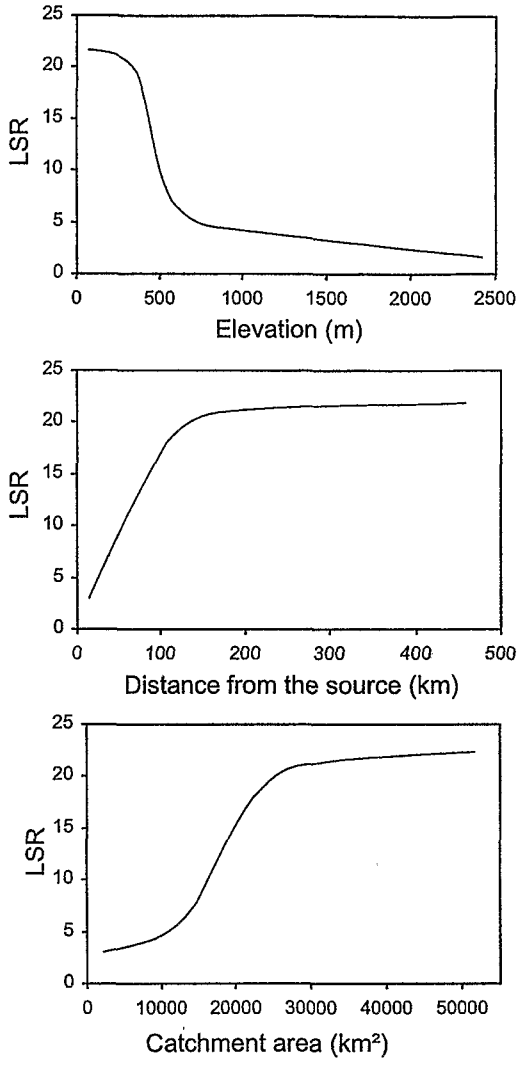

Figure 4. Sensitivity profiles of the three independent variables to the predicted values of local fish species richness (LSR).

4. Discussion

To our knowledge, it is the first time (in ecology), that a validation of a neural model is realized on a completely independent database and local species richness has been well fitted through ANN analysis by using three environmental characteristics. The ANN approach deserves therefore some attention in community ecology and biodiversity studies, where poor fitting of biological characteristics to conventional models (mostly multiple regression, MR) is often the rule.

Previous studies predicting fish species richness in different river systems (distributed across a wide range of latitudes) using logarithmic transformation of variables (regardless of the number of variables used) succeeded in explaining up to around 50 % of the total variation in richness [13, 15, 17], whereas the ANN method achieved, in our study, a much higher level (ca 70 %) with only three environmental variables.

A theoretical advantage of conventional models over ANNs is that their parameters provide information about the relative importance of the independent variables (although this is not true when composite variables are used). Nevertheless, the same results can be obtained by performing a sensitivity analysis of the ANN. To illustrate the explanatory variable importance inside the ANN, a procedure for partitioning the neural network connection weights in order to determine the relative importance of the various input variables has been proposed [28, 29]

and an algorithm allowing the visualization of the profiles of explanatory variables built [18, 23]. This characterization of the role of each variable in the ANN model clearly exhibits the non-linear processes according to its biological bases (*figure 4*).

Our results confirm the longitudinal change in local fish species richness along an upstream–downstream gradient. The fish species richness increases with increases in river size until middle reaches to level out in downstream areas. This relationship has been demonstrated in other river systems [8, 9, 11, 15] and contrasts with results obtained for the Seine basin where species richness declines in downstream areas [6, 10]. Thus, these findings confirm the hypothesis made by Oberdorff et al. [6] that the gradual decline in fish species richness observed for the lower Seine was a consequence of the loss of natural lotic habitats due to anthropogenic disturbances rather than a natural ecological pattern.

Considering only a few environmental variables (three, here) and a relatively simple neural network, one can explain the major part of the variability of fish species richness on a local scale. Thus, it would be now interesting, using this approach, to generate comparative studies with other large rivers of France, Europe, or elsewhere in the world to analyse factors controlling spatial variability in local fish species richness.

The ANN modelling approach introduced here is a fast and flexible way to incorporate multiple input parameters into one model. It is this ability to deal with multiple information sources that provides the main power of this approach, which results in a significant improvement in modelling over conventional approaches [30].

5. References

[1] Huet M., Profiles and biology of Western European streams as related to fish management, Trans. Am. Fish. Soc. 88 (1959) 155–163.

[2] Changeux T., Structure du peuplement piscicole à l'échelle d'un grand bassin européen: organisation longitudinale, influence de la pente et tendances régionales, Bull. Fr. Pêche Piscic. 337/338/339 (1995) 63–74.

[3] Verneaux J., Biotypologie de l'écosystème « eau courante ». Détermination approchée de l'appartenance typologique d'un peuplement ichtyologique, C. R. Acad. Sci. (III) 284 (1977) 675–678.

[4] Horwitz R.J., Temporal variability patterns and the distributional patterns of stream fishes, Ecol. Monogr. 48 (1978) 307–321.

[5] Beecher H.A., Dott E.R., Fernau R.F., Fish species richness and stream order in Washington State streams, Environ. Biol. Fish. 22 (1988) 193–209.

[6] Oberdorff T., Guilbert E., Lucchetta J.C., Patterns of fish species richness in the Seine River basin, France, Hydrobiologia 259 (1993) 157–167.

[7] Belliard J., Boët P., Tales E., Regional and longitudinal patterns of fish community structure in the Seine River basin, France, Environ. Biol. Fish. 50 (1997) 133–147.

[8] Mahon R., Divergent structure in fish taxoscenes of north temperate streams, Can. J. Fish. Aquat. Sci. 41 (1984) 330–350.

[9] Balon E.K., Crawford S.S., Lelek A., Fish communities of the upper Danube River (Germany, Austria) prior to the new Rhein-Main-Donau connexion, Environ. Biol. Fish. 15 (1986) 243–271.

[10] Belliard J., Le peuplement ichtyologique du bassin de la Seine. Rôle et signification des échelles temporelles et spatiales, thèse, université Paris-VI, 1994, 197 p.

[11] Changeux T., Structure des peuplements de poissons à l'échelle du bassin rhodanien. Approche régionale et organisation longitudinal. Exploitation des captures par pêche aux engins, thèse, université Claude-Bernard-Lyon-I, 1994, 242 p.

[12] Gorman O.T., Karr J.R., Habitat structure and stream fish communities, Ecology 69 (1978) 1239–1250.

[13] Angermeier P.L., Schlosser I.J., Species-area relationship for stream fishes, Ecology 70 (1989) 1450–1462.

[14] Rahel F.J., Hubert W.A., Fish assemblages and habitat gradients in a rocky mountain-great plains stream: biotic zonation and additive patterns of community change, Trans. Am. Fish. Soc. 120 (1991) 319–332.

[15] Pont D., Belliard J., Boët P., Changeux T., Oberdorff T., Ombredane D., Analyse de la richesse spécifique de quatre ensembles hydrographiques français, Bull. Fr. Pêche Piscic. 337/338/339 (1995) 75–81.

[16] Lim P., Belaud A., Labat R., Peuplement piscicole de la Garonne entre Saint-Gaudens et Agen, Ichtyophysiologica Acta 9 (1985) 187–201.

[17] Hugueny B., Richesse des peuplements de poissons dans le Niandan (haut Niger, Afrique) en fonction de la taille de la rivière et de la diversité du milieu, Rev. Hydrobiol. Trop. 23 (1990) 351–364.

[18] Lek S, Belaud A., Baran P., Dimopoulos I., Delacoste M., Role of some environmental variables in trout abundance models using neural networks, Aquat. Living Resour. 9 (1996) 23–29.

[19] Mastrorillo S., Lek S., Dauba F., Predicting the abundance of minnow *Phoxinus phoxinus* (Cyprinidae) in the River Ariège (France) using artificial neural network, Aquat. Living Resour. 10 (1997) 127–134.

[20] Mastrorillo S., Lek S., Dauba F., Belaud A., The use of artificial neural networks to predict the presence of small-bodied fish in a river, Freshwater Biol. 38 (1997) 237–246.

[21] Guégan J.F., Lek S., Oberdorff T., Energy availability and habitat heterogeneity predict global riverine fish diversity, Nature 391 (1998) 382–384.

[22] Boët P., Fuhs T., Les réseaux de neurones pour prédire la biodiversité des poissons en eau courante, Ingénieries-EAT 4 (1995) 5–14.

[23] Lek S., Delacoste M., Baran P., Dimopoulos I., Lauga J., Aulanier S., Application of neural networks to modelling nonlinear relationships in ecology, Ecol. Model. 90 (1996) 39–52.

[24] Décamps H., Fortuné M., Gazelle F., Pautou G., Historical influence of man on the riparian dynamics of a fluvial landscape, Landscape Ecology 1 (1988) 163–173.

[25] Dynesius M., Nilsonn C., Fragmentation and flow regulation of river systems in the northern third of the world, Science 266 (1994) 753–762.

[26] Rumelhart D.E., Hinton G.E., Williams R.J., Learning representations by back-propagating error, Nature 323 (1986) 533–536.

[27] Abdi H., A neural network primer, J. Biol. Sys. 2 (1994) 247–281.

[28] Garson G.D., Interpreting neural network connection weights, A.I. Expert. 6 (1991) 47–51.

[29] Goh A.T.C., Back–propagation neural networks for modeling complex systems, A.I. Eng. 9 (1995) 143–151.

[30] Zhang Q., Stanley S.J., Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling, Water Res. 31 (1997) 2340–2350.