Stream fish assemblages and basin land cover in a river network

Young-Seuk Park a,⁎, Gaël Grenouillet b, Benjamin Esperance b, Sovan Lek b

a Department of Biology, Kyung Hee University, Dongdaemun-gu, Seoul 130-701, Republic of Korea
b Laboratoire Dynamique de la Biodiversité (LADYBIO), UMR CNRS-Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 4, France

Available online 19 April 2006

Abstract

This study focused on characterizing fish assemblages in the Adour–Garonne basin and identifying the relative influences of landscape-scale features on observed patterns in stream fish assemblages. Two different artificial neural network algorithms were used: a self-organizing map (SOM) and a multilayer perceptron (MLP). A SOM was applied to determine fish assemblage types, and a MLP was used to predict the fish assemblage types defined by the SOM. Thirty four species were collected at 191 sampling sites in a major river-system, the Adour–Garonne basin, and topographical factors, namely altitude, distance from source and surface area of drainage basin were measured. Using GIS, land cover types (agricultural land, forests and urbanized artificial surface) were calculated for each site and expressed as percentage of the surface area of basin. These variables were introduced to the MLP and factorial discriminant analysis for the prediction of assemblage types. As a result, the SOM distinguished three fish assemblage types according to the differences of species composition, and the assemblage types were better predicted with landscape-scale features by MLP than discriminant analysis. The percentages of agricultural land and the surface area of a basin showed the greatest influence on assemblage types 1 and 2, and distance from source was the most important factor to determine assemblage type 3.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Fish; Assemblages; River; Basin; Land cover; ANN; GIS

1. Introduction

Previous studies have provided evidence that measurements of both local and landscape variables are useful predictors of biotic variables (Allan and Johnson, 1997; Brosse et al., 2003). Although the question of the relative importance of local versus landscape control of biological assemblages has a long history, relating biodiversity and land-use intensity has only recently received increased attention. During the last decade, studies demonstrated that changes in land use were one of the main driving forces behind biodiversity changes at the regional scale (van Diggelen et al., 2005). Some authors reported that land use throughout a river basin was a better predictor of habitat quality and biotic integrity at a site than either riparian corridor or reach surveys (e.g., Allan et al., 1997). It is thus generally accepted that quantification of land use is a valuable indicator of the state of ecosystems (Meyer and Turner, 1994).

In running water systems, numerous studies have documented the relationships between human activities at the landscape scale and the ecological integrity of a stream (Allan, 2004). Stream biodiversity is strongly influenced by land use within the surrounding valley, and previous studies pointed out that land use can impact the condition of a stream condition by both direct
influences (e.g., habitat, water quality) and interactions with other anthropogenic drivers (e.g., climate change, invasive species, dams). To date, most landscape-scale studies of the influence of land use on stream biodiversity have focused on basins dominated by agriculture or urban land. Many studies have reported that agricultural land use degrades water quality, impacts both riparian and stream habitat quality, and alters flows (Allan, 2004). Thus, declines in biodiversity have been documented for macroinvertebrate and fish assemblages in basins with a greater cover of agricultural land (Genito et al., 2002; Wang et al., 1997; Harding et al., 1998). Similarly, declines in biodiversity have also been reported as the extent of urban land increases, and despite a low percentage of the total surface area of a basin, urban land use can exert major influence on water quality, habitat and biological assemblages (Allan, 2004).

In fish ecology, the structure of local fish assemblages is dependant on numerous interacting factors, and many studies have addressed the question of the relative importance of local- versus regional-scale ecological processes (e.g., Angermeier and Winston, 1998). Although it has been widely recognized that local and regional factors can both be important in structuring local assemblages (Matthews, 1998), some authors have noticed that the two perspectives are still often considered independently (Fangliang et al., 2005) and that the majority of studies have failed to acknowledge differing scales associated with these factors (Van Zyll de Jong and Cowx, 2005). For stream fish populations, the effects of land use are indirect and can act through stream hydrology (e.g., flow patterns), geomorphology, water chemistry, sediment transport, and riparian vegetation (Jowett et al., 1996; Wang et al., 2001). For example, land use can influence substrate composition and substrate stability, determining thus the amount of suitable fish habitat. Increase in urban land use has been shown to increase precipitation runoff volume, modifying the frequency and magnitude of floods, which in turn make the channel less stable (Wang et al., 2001). Thus, the influences of in-stream habitat and land use appear to be inter-related (Hanchet, 1990), and as in-stream habitat varies between land use types, separating their relative effects remains difficult. Therefore, quantifying and understanding how land use affects stream processes and biological assemblages are crucial questions for determining how to minimise human impacts on stream ecosystems (Roy et al., 2003).

The aim of this study was to examine the relative influence of landscape-scale features on stream fish distribution. Here we focused on fish assemblage patterns in the Adour–Garonne basin. For this purpose we used fish species lists (i.e., presence–absence). Two approaches (i.e., Artificial Neural Networks and Discriminant Analysis) were used to investigate fish–habitat relationships. The objectives of this study were to: (1) characterise fish assemblages in the Adour–Garonne basin, (2) identify the relative influences of landscape-scale features on observed patterns in stream fish assemblages, and (3) discuss the utility of land cover variables as a tool in a management perspective, in particular to predict the effects of regional changes (i.e., land use modifications) on running water systems. To our knowledge, this study is the first to address the relative effects of both geomorphologic and land use features on stream fish assemblages in this river basin. This work contributes to the modelling assessment of the factors and processes controlling freshwater ecology within the Eurolimpacs project (Wade, 2006–this volume).

2. Materials and methods

2.1. Dataset

2.1.1. Study area

The study area of the Adour–Garonne river network covers approximately 116 000 km² (Fig. 1). The network has a total length of about 120 000 km and presents a...
wide range of altitudes (high mountains to plain and coastal areas) and geological substrates (calcareous, sedimentary, sandstone, crystalline and volcanic). The Adour–Garonne consists of 7 main sub-basins related to the following rivers: Charente, Dordogne, Lot, Tarn-Aveyron, Garonne, Adour and coastal streams. The main channel, the Garonne, runs 580 km from the Pyrenean mountains to the Gironde estuary at the Atlantic coast. Its tributaries which have their sources in the Massif Central plateau and the Pyrenees, which are barriers of dispersal for fish, have a lower number of fish species (45, including estuarine and migratory species) in comparison with other basins in France (Keith and Allardi, 2001).

2.1.2. Fish datasets

Fish assemblage data from 191 sample sites were collected between 1986 and 1996 and obtained from the fish database of the Aquatic Environment Team, School of Agronomy at Toulouse (ENSAT) and from the French Fisheries Council (CSP). Electrofishing surveys were done either by wading in shallow areas or by boat in the deeper reaches. In the case of wider and deeper rivers, gill-netting was used in still waters and both gill- and drift-netting for running waters. This combination of methods allows an effective assessment of fish diversity in rivers (Seegert, 2000). However, as abundance measures reflect collection intensity (Angermeier and Smogor, 2005), only presence–absence data were considered to remove sampling bias as recommended by Hughes and Gammon (1987). Indeed, many studies assuming large spatial and temporal scales use species presence–absence as the level of data resolution due to the difficulties in obtaining reliable estimates of relative rank abundance (Jackson et al., 2001).

All 191 sites were not sampled in every year due to financial constraints. For the analyses, a large spatio-temporal scale was therefore assumed: a 10-year span as the time unit, and the whole Adour–Garonne basin as the spatial unit. Moreover, local repeated surveys at some sampling sites showed that both environmental characteristics and fish assemblages did not vary significantly over the 10-year sampling period. Thus, the time dimension could be factored out, pooling all observations together in a similar way, as described by Sipponen and Muotka (1996). We considered this database, nevertheless, a reliable representation of fish fauna in this area, according to our field experience and with the information of fish atlases (Bruslé and Quignard, 2001; Keith and Allardi, 2001). In the dataset of 191 samples, 34 species were identified, and used in the modelling procedure.

2.1.3. Environmental factors

Among the environmental factors determining the spatial distribution of stream fish species, the upstream–downstream gradient is one of the key factors influencing stream fish assemblages. This longitudinal pattern reflects an increase in fish species richness with increasing stream size (Matthews, 1998). Although stream fish assemblages respond to many physical and chemical factors, factors such as flow velocity, macrophyte abundance or water quality were not available for all sampling sites. Therefore, we focused on environmental descriptors reflecting the position of each sampling site along the upstream–downstream gradient and three environmental descriptors were chosen according to their importance in explaining topography of sampling site: altitude (ALT), distance from source (DIST) and the surface area of the drainage basin (SURF). Additionally, the proportion (%) of different land covers in the basin surface area of each sampling site was extracted from the CORINE Land Cover map (CLC; http://www.ifen.fr) through MAPINFO 5.0. CLC is a map of the European environmental landscape intended for use by policy makers as well as scientists. Based on interpretation of satellite images, CLC provides comparable digital maps of land cover for each country for much of Europe. It is based on a simple 3-level hierarchy classification system consisting of 44 land cover classes. The minimum area digitised was 25 ha of homogeneous cover of one single class. We used three classes in the first level of CLC categories: urbanized artificial surface (ART), agricultural land (AGR), and forest area (FOR). Percentages of these three classes were used in the modelling procedure, whereas other classes such as wetlands and water bodies were not included. Artificial surfaces in CLC include four subcategories: 1) urban fabric, 2) industrial, commercial and transport units, 3) mine, dump and construction sites, and 4) artificial non-agricultural vegetated areas. Some variables such as agricultural land and forest are closely related to each other showing significant correlation coefficients (Table 1).

2.2. Modelling procedure

Explaining the variation of ecological data can be considered in two steps: ordination methods to summarize the variability of the data as a first step and exploration of possible relationships between biological and environment variables as a second step (Jongman et al., 1995). From this point of view, to pattern and predict fish communities, we used two different artificial neural networks (ANNs): a self-organizing map (SOM) which...
is an unsupervised neural network (Kohonen, 1982, 2001) and a multilayer perceptron (MLP) with a backpropagation learning algorithm which is a supervised neural network (Rumelhart et al., 1986). The SOM approximates the probability density function of the input data and is a method for clustering, visualization and abstraction, the purpose of which is to visualize the dataset in a more usable form (Kohonen, 2001). Additionally, the SOM averages the input dataset in weight vectors through the learning process and thus removes noise (Vesanto et al., 1998). The MLP is a mathematical algorithm to extract relationships between explanatory and response variables, and offers an effective approach to the computation of the model coefficients (Kung, 1993). In this study, first, the SOM as an ordination method was applied to distinguish fish communities according to the similarity of fish species and to summarize the variability of the biological data. Thus, sampling sites were arranged on plots which summarized the spatial variability of the biological features and independently, the underlying environmental conditions. At the second step of the analysis, the MLP was used to predict the community types identified by the SOM.

### 2.2.1. Patterning samples with SOM

The SOM consisted of two layers: an input layer formed by a set of units (or neurons which were computational units), and an output layer formed by units arranged in a two-dimensional grid. In this study, each input unit accounted for the abundance of one species, which led to an input layer made of 34 neurons (for each of 34 species). The output layer was made of output units in a hexagonal lattice (i.e. in this study 72 units in a grid of 12×6 cells) which provided better visualization. The map size (number of output units) of the SOM was critical for accommodating hierarchical levels in community classification. We trained the SOM with different map sizes, and chose the optimum map size (72=12×6) based on the minimum values of quantization and topographic errors. A hexagonal lattice was preferred because it does not favour horizontal or vertical directions (Kohonen, 2001). Formally, the SOM algorithm maps a set of input vectors (i.e. samples in this study) onto a set of output vectors, according to the characteristics of the input vector components (i.e., species in this study). It can be interpreted as a non-linear projection of the high-dimensional input data onto an output array of neurons. Each output unit has a vector of coefficients associated with the input data. The coefficient vector is referred to as a weight (or connection intensity) vector \( W \) between input and output layers. The weights play an important role in the propagation of the signal through the model: they establish a link between the input units and their associated output units.

The algorithm can be described as follows: when an input vector \( X \) (i.e. the presence/absence of 34 species in each sample) is presented to the network, the neurons in the output layer compete with each other and the winner (whose weight has the minimum distance from the input vector) is chosen. The winner and its neighbours predefined in the algorithm update their weight vectors according to the SOM learning rules as follows:

\[
\begin{align*}
\mathbf{w}_i(t+1) &= \mathbf{w}_i(t) + \alpha(t) \cdot h_{jc}(t)[x_j(t) - \mathbf{w}_j(t)], \\
\end{align*}
\]

where \( \mathbf{w}_i(t) \) is a weight between a neuron \( i \) in the input layer and a neuron \( j \) in the output layer at iteration time \( t \), \( \alpha(t) \) is a learning rate factor which is a decreasing function of the iteration time \( t \), and \( h_{jc}(t) \) is a neighbourhood function (a smoothing kernel defined over the lattice points) that defines the size of neighbourhood of the winner neuron \( c \) to be updated during the learning process. The output is scaled in the range 0 to 1 and the explicit form of \( h_{jc}(t) \) is as follows:

\[
\begin{align*}
h_{jc}(t) &= \exp\left(-\frac{||x_j - r_c||^2}{2\sigma^2(t)}\right),
\end{align*}
\]

where the parameter \( \sigma(t) \) is a monotonically decreasing function of iteration time \( t \) defining the width of the
kernel and \( |r_j - r_c| \) is the distance in the output map between the winner neuron \( (c) \) and its neighbor neuron \( (j) \). This learning process is continued until a stopping criterion is met, usually, when weight vectors stabilize or when a number of iterations (6000 iterations in this study) are completed. This learning process results in training the network to pattern the input vectors and preserves the connection intensities in the weight vectors.

On the trained SOM map, it is difficult to distinguish subsets because there are still no boundaries between possible clusters. Therefore, it is necessary to subdivide the map into different groups according to the similarity of the weight vectors of the neurons. To define clusters between SOM units, we used a hierarchical cluster analysis with a Ward’s linkage method (Legendre and Legendre, 1998) by calculating the Euclidean distance between weight vectors of each SOM unit. The SOM is an effective learning algorithm for the visualization and abstraction of high-dimensional data in low dimensions. It converts the non-linear statistical relationships between high-dimensional data into simple geometric relationships of their virtual units on a low-dimensional display, usually a two-dimensional lattice (Kohonen, 2001). However, the visualization of the trained SOM only reveals qualitative information. By implementing two-level classifications, the distances between the groups in the SOM map can be efficiently defined through hierarchical levels (Park et al., 2004).

During the learning process of the SOM, map units that are topographically close in the array will activate each other to learn something from the same input vector. This results in a smoothing effect on the weight vectors of neurons (Kohonen, 2001). These weight vectors tend to approximate the probability density function of the input vector. Therefore, the visualization of elements (i.e. fish species) of these vectors is convenient to understand the contribution of each input variable with respect to the clusters on the trained SOM. To analyse the contribution of variables to cluster structures of the trained SOM, the weight vector of each input variable (species) calculated during the training process was visualised in each neuron on the trained SOM map in grey scale.

To understand relationships between biological and environmental variables, we calculated mean values of each environmental variable in each cluster defined in the SOM map. Differences between clusters defined through the SOM were analysed using multi-response permutation procedures (MRPP) (PC-ORD ver. 4.25), analysis of variance (ANOVA), and Tukey’s Honestly Significant Differences (HSD) test for multiple comparisons of variables. Details of the application of SOMs in ecological studies can be found in Park et al. (2003).

2.2.2. Prediction of patterns with MLP

The prediction of assemblage types with their 6 environmental variables was done using the MLP which is designed to minimize the mean square error between the computed output of the network and the desired output; for a detailed description of the learning rules of the MLP see Rumelhart et al. (1986), Kung (1993), and Lek and Guégan (2000). The network usually consists of three layers: an input layer, one or more hidden layers and an output layer. Each layer is composed of neurons, which are the computational units of the MLP. It requires input vectors in the input layer, as well as target (or desired) values in the output layer corresponding to each input vector. The input layer contains neurons for the independent variables.

We used 6 input neurons for 6 environmental variables in this study. The output layer was composed of three neurons, one for each cluster, and the highest output value of the three nodes determined the cluster identity. The MLP model was trained with different number of hidden neurons (from 2 to 15), and large number of hidden neurons showed relatively better prediction results. We chose 7 hidden neurons of a single hidden layer to avoid an over-fitting problem. In the MLP, signals are propagated from the input layer through the hidden layer to the output layer via the network connections. During the training phase, a comparison was made between the output values calculated by the MLP and the expected values, and the connection weights were modified to minimise the error of the response: the difference between the expected and calculated output values.

We used a leave-one-out procedure (Kohavi, 1995) to validate the MLP model with 191 samples. The network was trained for 500 iterations. The performance of the MLP models was evaluated using the percentage of correct answer (i.e. percentage of correctly predicted samples) and Cohen’s Kappa statistics (Cohen, 1960).

After the learning process of the MLP model, a sensitivity analysis was carried out to evaluate the contribution of each input variable to the output values. There are several ways to perform the sensitivity analysis with MLP (Zurada et al., 1994; Dimopoulos et al., 1999). We used the partial derivatives (PaD) method because it is more coherent from a computational point of view (Dimopoulos et al., 1995, 1999). The PaD method presents the output of the MLP models with respect to the input to obtain the profile of the variations of the output for small changes of one input.
variable (Dimopoulos et al., 1995, 1999). The formula for the partial derivatives \( d_{ji} \) is:

\[
d_{ji} = S_j \sum_{h=1}^{nh} w_{ho} I_{ij} (1 - I_{ij}) w_{ih}
\]

where \( S_j \) is the derivative of \( j \)th output neuron with respect to its input, \( I_{ij} \) is the response of the \( h \)th hidden neuron, \( w_{ho} \) and \( w_{ih} \) are weights between \( o \)th output neuron and \( h \)th hidden neuron, and between the \( i \)th input neuron and the \( h \)th hidden neuron, respectively.

The learning process of the MLP was repeated 15 times to evaluate the variation of the contribution by randomly initializing weights of the MLP, and the contribution of each input variable was evaluated for each repetition.

If the partial derivative of a variable is negative then, for each parameter being analyzed, its influence tends to be negative on the determination of output variables (i.e. clusters in this study): occurrence probability of a particular cluster is lowered. Inversely, if the partial derivative is positive, the influence tends to be positive.
on the output variable. When the partial derivative is around zero, the influence is small. The relative contribution of input descriptors to the MLP output can be estimated as the sum of the squared partial derivatives (SSD) obtained for each input variable. The SSD values allow the classification of the variables according to their contribution to the output variable in the model, the input variable with the highest SSD value being the variable which most influences the output variable.

To compare the performance of prediction and evaluation of parameters of MLP, Discriminant Function Analysis (DFA) was conducted using R software (Ihaka and Gentleman, 1996). The basic idea underlying DFA consists of finding the linear combination of explanatory variables which best separate groups, based upon maximizing between group variance while minimizing within group variance. Standardized coefficients for each variable in each discriminant function represent the contribution of the respective variable to the discrimination between groups. DFA was used to determine which environmental variables discriminate between the groups previously defined by the hierarchical cluster analysis according to the similarity of fish species. Finally, we performed cross-validation to test the ability of the model to discriminate between groups, so that each observation was classified by the functions derived from all observations other than that observation (leave-one-out process).

3. Results

3.1. Patterning sampling sites

Through the learning process of the SOM, 191 sampling sites were patterned on the SOM map.

Fig. 4. Box plots showing occurrence probability (%) of each species in different clusters. Values were obtained from weight vectors of the trained SOM. – median, 25–75%, non-outlier range. Different colours are used simply to aid the reader in the identification of different species across the 3 clusters.
according to the similarity of their species occurrences. The hierarchical cluster analysis showed 2 large groups with different subgroups at different linkage distance (Fig. 2a). Based on ecological features of each clusters and the comparison of a dendrogram of the cluster analysis and U-matrix (Ultsch, 1993), we defined 3 clusters, which were significantly different in the species composition (MRPP, \( P < 0.001 \)). Overall, samples showing high species richness were located in clusters 2 and 3, whereas samples with low species richness were in cluster 1 (Fig. 2b). Samples of each cluster were presented on the geographical map, showing an upstream–downstream gradient (Fig. 1). Samples in cluster 1 were located mostly in the upstream areas, whereas samples in cluster 3 were in the downstream areas. Samples in cluster 2 corresponded to mid-stream areas.

### 3.2. Contribution of species in patterning

The contribution of species to community grouping was examined using connection intensities of the SOM. Through the learning process of the SOM, the approximate probability density function of input data was calculated as connection intensities. Fig. 3 displays example component planes of typical distribution patterns of species using the connection intensity of each species in the SOM units on a grey scale. Visualization of this connection is an efficient way to comprehensively understand the contribution of each species in SOM patterning. The connection intensity represents approximately the occurrence probability of each species in patterned samples. The SOM showed that the centres of
the distributions of most species were clearly associated with single clusters (Fig. 4). For example, brown trout (Salmo trutta fario) showed the highest occurrence probability in cluster 1, soiffe (Chondrostoma toxostoma) was most likely to occur in cluster 2, and Ictalurus melas in cluster 3. However, some species such as minnow (Phoxinus phoxinus) and gudgeon (Gobio gobio) were commonly observed in three clusters, and roach (Rutilus rutilus) was highly distributed in clusters 2 and 3. From the occurrence probability of each species, community assemblages could be defined for each cluster (hereafter, assemblage type). Species richness significantly (Pb0.001, Tukey’s HSD test) differed between assemblage types, showing an increase in species richness from type 1 to type 3 (Fig. 5).

3.3. Differences of environmental variables

Each variable significantly differed between clusters (Fig. 6). Altitude was highest in cluster 1, and lowest in cluster 3. In contrast, the surface area of the basin and the distance from the source showed similar patterns, with the highest values in cluster 3, and lowest ones in cluster 1. Among land cover variables, the proportion of forest area was highest in cluster 1 than in other clusters. In contrast, artificial surface and agricultural land showed lowest values in cluster 1. Therefore, these results confirmed that cluster 1 corresponded to mountain areas, whereas cluster 3 corresponded to downstream areas, reflecting the longitudinal gradient. Clusters 2 and 3 were not significantly different in percentage of land covers, whereas they were significantly different in topographical variables (altitude, surface area of basin, and distance from source).

3.4. Prediction of fish assemblage types

The MLP showed a high predictive capability with higher than 81% of predictions being correct: among 108 samples of assemblage type 1, on average about 100 samples were correctly predicted in each repetition (Table 2). Among the three assemblage types, type 2 was the least well predicted with about 67% of correct answers correct. Assemblage type 3 had 85% of predictions correct. Cohen’s Kappa was 0.76, indicating very high agreement between assemblage types predicted by the MLP and the corresponding expected types defined by the SOM. The prediction power (both % correct answer and

Table 2
Prediction capability of a multilayer perceptron (MLP) and discriminant function analysis for three assemblage types with 6 environmental variables

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Assemblage type</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Number of samples</td>
<td>108</td>
<td>37</td>
</tr>
<tr>
<td>Multilayer perceptron</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of correctly predicted samples</td>
<td>(2.48)</td>
<td>(3.16)</td>
</tr>
<tr>
<td>% correct answer</td>
<td>92.41</td>
<td>67.57</td>
</tr>
<tr>
<td>Cohen’s kappa</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Discriminant function analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of correctly predicted samples</td>
<td>79</td>
<td>26</td>
</tr>
<tr>
<td>% correct answer</td>
<td>73.15</td>
<td>70.27</td>
</tr>
<tr>
<td>Cohen’s kappa</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

The numbers in the parenthesis are standard deviations.

Fig. 7. Relative contribution of environmental variables to the determination of fish assemblage types. A partial derivatives method was applied for the sensitivity analysis. The same letters on the bars indicate no significant differences based on Tukey’s HSD test (P=0.05). Error bars indicate standard error.
Cohen’s Kappa) of the discriminant analysis was lower than that of the MLP model (Table 2).

3.5. Relative importance of environmental factors

The percentage cover of agricultural land was the most important factor in determining assemblage type 1, followed by the surface area of basin expressed as a percentage (Fig. 7). For assemblage type 2, the surface area of basin was the most important, followed by percentage of agricultural land. For assemblage type 3, distance from the source was the most important factor accounting for about 58% of the variation explained, followed by the surface area of basin and altitude, while land cover variables showed low contributions.

Partial derivatives of agricultural land and the surface area of basin were mostly negative (Fig. 8). In particular when percentage cover of agricultural land was large, the derivatives were scattered around zero values when the percentage cover of agricultural land was low. This implies that percentage cover of agricultural land and surface area of basin influenced negatively the determination of assemblage type 1, especially when agricultural land cover was higher than 40%. In contrast, for assemblage type 2, the surface area of basin contributed positively in its low range values (less than about 3000 km²), while a negative contribution was observed in the middle ranges (about 3000–15 000 km²). Percentage cover of agricultural land showed mainly negative effects on the determination of assemblage type 1.

Fig. 8. Partial derivatives of first two variables contributing to each assemblage type (indicated by numbers in parentheses). Logarithmic scale was used for surface area of basin for helping to better discern the patterns.

Partial derivatives of agricultural land and the surface area of basin were mostly negative (Fig. 8). In particular when percentage cover of agricultural land was large, the derivatives were scattered around zero values when the percentage cover of agricultural land was low. This implies that percentage cover of agricultural land and surface area of basin influenced negatively the determination of assemblage type 1, especially when agricultural land cover was higher than 40%. In contrast, for assemblage type 2, the surface area of basin contributed positively in its low range values (less than about 3000 km²), while a negative contribution was observed in the middle ranges (about 3000–15 000 km²). Percentage cover of agricultural land showed mainly negative effects on the determination of assemblage type 1.

Fig. 8. Partial derivatives of first two variables contributing to each assemblage type (indicated by numbers in parentheses). Logarithmic scale was used for surface area of basin for helping to better discern the patterns.

Fig. 9. Discriminant analysis. Given are (a) eigenvalues, (b) factorial map of environmental variables and (c) factorial map of sites grouped by assemblage types (indicated by numbers in squares). DIS; distance from the source, ALT; altitude, SUR; surface area of basin, ART; artificial surface, AGR; agricultural land, FOR; forest.
type 2 at the low values (0–30% of the range), while positive influence was observed at high values of agricultural land cover (50–90% of the range). Assemblage type 3 was mainly determined by distance from source and the surface area of basin, but partial derivatives showed that the influence of distance from source was weak at higher than 250 km.

According to discriminant analysis (Fig. 9), assemblage types were significantly separated ($P=0.001$, randomization test). The three assemblage types were ordered along the first axis of the analysis, which accounted for 87% of total inertia, but assemblage type 2 clearly overlapped the two other types. Along the first axis, distance from source and surface area of basin were the two most contributing variables separating assemblage types. Along the second axis (13% of total inertia), agricultural land and forest area, were the two variables contributing most discriminatory power (Fig. 9).

4. Discussion

The main objectives of this study were to characterize fish assemblages in the Adour–Garonne basin, and to identify the relative influences of landscape-scale features on observed patterns in stream fish assemblages. Sampling sites were classified through the learning process of the SOM, according to fish assemblage similarities and mainly reflected the longitudinal (i.e., upstream–downstream) gradient. Based on this classification, three fish assemblage types were defined: mountain areas (type 1) characterized by brown trout ($S. trutta fario$) and minnow ($P. phoxinus$), mid-stream areas (type 2) characterized by soiffe ($C. toxostoma$) and barbell ($Barbus barbus$), and downstream areas (type 3) characterized by belica ($Leucaspius delineatus$) and black bullhead ($I. melas$). Biotic river zonation has been developed based on the longitudinal change in fish assemblage structure along the upstream–downstream gradient, (Thienemann, 1925; Huet, 1949, 1954). Huet (1949) zonation consists of four zones, beginning with the headwater and moving to the lowlands (i.e., the trout zone, the grayling zone, the barbel zone, and the bream zone). This classification is well observed in our study, although we defined only 3 types in the Adour–Garonne basin. Assemblage type 1 in our study mainly corresponds to the trout zone, type 2 corresponds to the grayling and barbel zones, and type 3 corresponds to the bream zone. Our results are consistent with previous studies, such as Rahel and Hubert (1991), which suggest that both the grayling and barbel zones would represent a transition between the trout and the bream zones. Although grayling is not a native species in the Adour–Garonne basin, fish often associated with it like brown trout, stone loach, gudgeon, and minnow (Mastrorillo et al., 1998; Reyjol et al., 2001) correspond to our assemblage type 2. Other studies have shown that fish assemblages from mountain areas and river mouth areas differ in species composition, while assemblages in transitional areas might represent a gradual transition of species (Naiman et al., 1988; Lyons, 1989, 1996; Paller, 1994; Aguilar-Ibarra et al., 2005).

In this study, fish assemblage types defined by the SOM and subsequent analysis were well predicted with 6 environmental variables through MLP models, showing 81%, 67% and 88% of correct prediction for types 1, 2, and 3, respectively. The prediction performance of the MLP model was better than that of the discriminant function analysis which showed 72% of correct prediction. Through the sensitivity analysis of the MLP model, the contribution of environmental variables to the prediction of assemblage types could be evaluated. For assemblage types 1 and 2, percentages of agricultural land and surface area of basin were the most important factors, whereas distance from the source was the most important factors for assemblage type 3.

Rivers are characterised by environmental gradients aligned longitudinally within the basin from source to mouth. In these systems, biological assemblages are organized longitudinally and influenced by the environment around the river system. Generally species richness increases according to the longitudinal changes from the source to the river mouth (e.g., measured by the river width, the distance from the source, the stream order, and the size of the watershed). Such a pattern was also observed in this study, with higher species richness in assemblage type 3 characterized by a large surface area of the basin and lower richness in type 1 characterised by mountain areas.

Percentage cover of artificial surface, agricultural land, and forest in the basin of sampling sites were significantly different between clusters (assemblage types) defined in the SOM, reflecting geographical distribution of sampling sites. For instance, percentage of forests was highest in cluster 1 showing high altitude and low surface area of basin, while proportions of agricultural land and artificial surface were highest in clusters 2 and 3.

Angermeier and Winston (1998) examined the relative influence of local versus regional factors on the local diversity of stream fish assemblages, and found regional factors to be most important. More recently, Pont et al. (2005) investigated riverine fish species responses to environmental constraints and tested the relative importance of local and regional factors in
species occurrence patterns. They showed that factors that operate at a regional scale (temperature and basin unit) have the highest explanatory power for rithronic fish species (e.g., brown trout and bullhead). However, the responses of these species differed from those of species inhabiting larger rivers, and these authors clearly illustrated that the sensitivity to local and regional scale processes is species specific, rejecting thus the hypothesis of a common species response to the environmental constraints. In this study, we tested the effects of both geomorphologic and land-use variables on fish distribution. One of the more interesting results is that factors influencing fish distribution varied along the longitudinal gradient. Our results clearly demonstrated the major importance of geomorphological variables for mid- and downstream fish assemblages (e.g., surface area of basin and distance from source, respectively), whereas land-use variables (e.g., percentage cover of agricultural land) were of major importance for upstream fish assemblages. These results are consistent with previous studies addressing the relative influence of local versus regional factors on fish distribution. As the responses of fish species to their environmental constraints are species specific (Pont et al., 2005), further work is needed to better identify the relative influence of each factor on species distribution. For the environmental variables considered, we focused on landscape-scale features but many other factors are known to influence stream fish distribution (reviewed in Matthews, 1998). Among these factors, those describing hydraulic conditions, stream morphology (e.g., slope and width), or thermal climate are important determinants of fish species distribution and could be added to improve the quality of our results.

Land use has been shown to be a determinant of the distribution of aquatic organisms which agrees with other studies (Carr, 1986; Lott, 1999). Some authors (e.g., Roy et al., 2003) showed that land-use variables correlated with a number of reach scale geomorphic variables and water chemistry variables, and that biotic indices based on macroinvertebrates were better predicted by reach scale variables than land-use variables. Among land-use variables tested in this study, only agricultural land appeared to be related to stream fish assemblages using ANN, whereas discriminant analysis also identified percentage of forests as an important correlate. These results are coherent with numerous studies which have documented declines in water quality, habitat, and biological assemblages as the extent of agricultural land increases. As agricultural land use increases, inputs of sediments, nutrients and pesticides increase, resulting in decline in habitat heterogeneity and water quality, increase in algal production, and alteration of flow regimes (reviewed in Allan, 2004). In fish studies, such impacts have been observed at both community (Rowe et al., 1999) and population (Hanchet, 1990) levels. For example, riparian forest removal has been shown to be associated with decreased abundance of benthic species, which were replaced by sediment-tolerant species (Dale Jones et al., 1999). Similarly, numerous studies have underlined that salmonid populations could be strongly impacted by agricultural land-use driven alterations (e.g. Pess et al., 2002).

Recent studies have pointed out that much more investigations across multiple spatial scales are required to understand whether and how processes might be hierarchically structured (Allan and Johnson, 1997). As land use features are potentially subject to modifications in the context of climate change, such studies within a landscape framework have important management implications. Although changes in biodiversity related to climate change have received an increasing interest, fewer studies documented the potential effects of other human-caused modifications. Previous studies have demonstrated that land-use changes could interact with climate change (Dale, 1997; Benning et al., 2002). These studies have demonstrated the difficulties inherent in estimating the relative influences of both climate and land-use changes, as climate change operates on a regional scale, while land use change operates at the local and national scales. An additional difficulty lies in the fact that land use change is related to climate change as both a causal factor and a major way in which the effects of climate change are expressed (Dale, 1997). Thus, making projections of future states of land use requires improved knowledge about (1) the factors that determine land use, and (2) how these factors interact to drive land use change. Such additional research is needed to use available climate change scenarios to project future patterns of land use and future rates of land use change. In this study, agricultural land use has been shown to be decisive, at least for upstream fish assemblages. Direct climate effects on agriculture are probably the best known, with changing temperature and rainfall patterns leading to changes in crop yield and spatial shifts of agricultural potential (see Dale, 1997). That would have severe implications for policy makers, and taking into account land use change scenarios appears then to be crucial for further understanding the potential impact of climate change on stream fish assemblages. Using projected changes in climate, previous studies have focused on changes in the thermal habitat likely to affect both fish spatial distribution and fish assemblage composition. According to predictions of future climate, we can expect some new species to
become established, while other species will likely decline due to direct or indirect effects related to climate change (e.g., Jackson and Mandrak, 2002). However, less attention has been paid to the potential effects of land-use changes and further studies are needed to increase our ability to predict the impacts of these changes on stream fish assemblages.

5. Conclusion

In summary, this study characterized fish assemblages in the Adour–Garonne basin and identified the relative influences of landscape-scale features on observed patterns in stream fish assemblages using two different artificial neural network algorithms, a self-organizing map (SOM) and a multilayer perceptron (MLP). The SOM distinguished three fish assemblage types reflecting the longitudinal gradient in the Adour–Garonne basin, and the assemblage types were better predicted with landscape-scale features by the MLP than the discriminant function analysis.

Acknowledgments

This work was supported by the EU project Euro-impacts (contract number GOEC-CT-2003-505540). We are grateful to Dr. A. Wade and two anonymous reviewers for their insightful comments on the early draft of this manuscript.

References


Kohavi R. A study of cross-validation and bootstrap for estimation and model selection. Proceedings of the 14th international joint