

Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages

S. Brosse *, J.L. Giraudel, S. Lek

CESAC, CNRS UMR 5576, Bat 4R3, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France

Abstract

Kohonen self-organizing maps (SOM) belong to the non-supervised artificial neural network modelling methods. It typically displays a high dimensional data set in a lower dimensional space. In this way, that method can be considered as a non-linear surrogate to the principal component analysis (PCA). In order to test the efficiency of SOM on complex ecological data gathered in the natural environment, we made a comparison between PCA and SOM capabilities to analyse the spatial occupancy of several European freshwater fish species in the littoral zone of a large French lake. The same data matrix consisting of 710 samples and 15 species was analysed using PCA and SOM. Both methods provided insights on the major trends in fish spatial occupancy. However, a more detailed analysis showed that only SOM was able to reliably visualise the entire fish assemblage in a two dimensional space (i.e. both dominant and scarce species). On the contrary PCA provided irrelevant ecological information for some species. These drawbacks were afforded to data heterogeneity, scarce species being poorly represented on the PCA plane. These results led us to conclude that SOM constitute a more reliable data representation method than PCA when complex ecological data sets are used. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Artificial neural networks; Kohonen self-organizing map; Lake; Fish assemblage; Principal component analysis

1. Introduction

Ecological applications of multivariate statistics have expanded tremendously during the last two decades (Gauch, 1982; Legendre and Legendre, 1998). Among these methods, the principal component analysis (PCA) is now used routinely by ecologists (Townsend et al., 1997; Grossman et al., 1998; Brosse et al., 1999a; Lamouroux et al., 1999). It is known as able to simplify large data

sets with reasonable loss of information and to assess intercorrelation among variables of interest (Grossman et al., 1991). However, the information given by PCA techniques suffers from some drawbacks in that the relationships between variables in environmental sciences are often non-linear (James and McCulloch, 1990), while the methods used are based on linear principles. Transformation of non-linear variables by logarithmic, power or exponential functions can appreciably improve the results, but have often failed to fit the data (Lek et al., 1996; Pennington, 1996; Brosse et al., 1999b). In the same way, ecologically relevant, but unusual observations,

* Corresponding author. Tel.: +33-5-61558687; fax: +33-5-61556096.

E-mail address: brosse@cict.fr (S. Brosse).

are frequently deleted from the data sets to reduce data heterogeneity (Gauch 1982; Copp et al., 1994). Although these deletions satisfy statistical assumptions, they are likely to bias the ecological interpretation of the results (Fore et al., 1996). To overcome these difficulties, the artificial neural networks which are known to be efficient in dealing with heterogeneous data sets should constitute a relevant alternative tool to traditional statistical methods (Lek et al., 1996; Lek and Guégan, 2000).

The self organizing map (SOM) algorithm, a non-supervised neural networks method, performs the same task as PCA (Kohonen, 1995). The main usefulness of this type of approach is to get an objective image of the populations assemblage with results uninfluenced by our knowledge of the samples and of the environmental features. Although SOM have scarcely been used in ecology, successful results were obtained by Aurelle et al. (1999) and Giraudel et al. (2000) for classification of fish populations on the basis of genetic data and by Chon et al. (1996) for communities patternizing. However, SOM and PCA capabilities to visualise data tendencies have only been compared using a well known reference data set (Giraudel et al. submitted) composed of few samples (10 samples and eight species). Although this data was already used by Ludwig and Reynolds (1987) and Chon et al. (1996) as an example for linear and non-linear statistical methods validation and explanation, it can not be considered as representative of the complexity of most experimental data usually gathered by ecologists.

In this work, we compared SOM and PCA capabilities to visualise in a two-dimensional space a complex data basis (710 samples and 15 species) composed of fish abundances in a large lake. The fish population assemblage obtained using the two methods were compared and discussed according to current ecological knowledge.

2. Material and methods

2.1. Study site and sampling

Studies were carried out on lake Pareloup. It is located in the south-west of France, near the city

of Rodez. It covers a total surface area of 1250 ha for a volume of approximately $168 \times 10^6 \text{ m}^3$. Fish sampling was performed in a restricted littoral zone of the lake using point abundance sampling by electrofishing (Nelva et al., 1979) to evaluate the spatial occupancy of the fish populations. Sampling was performed in July 1998 giving rise to 710 samples. It is during this month that fish species richness is maximal in the littoral zone (Brosse, 1999). For each sample, fish were counted and determined at the species level. When fish larvae and juveniles were collected, they were preserved in four-percent formaldehyde solution and then identified and numbered at the laboratory. For each fish species, individuals recorded were divided into two populations: young of the year (0+) and older fish called adults. This separation was done to avoid biases induced by the different spatial occupancy of 0+ and adults due to individuals habitat and feeding characteristics depending on size and age, as underlined by Persson and Greenberg (1990). For reading convenience, throughout this paper, we refer to analysis of 15 fish species, although two of the fish groups are in fact different age classes of the same species.

2.2. Data analyses

In order to extract the structure of the high-dimensional data formed by the 710 sample units (SUs) of abundance of the 15 species, two methods were used: firstly, a commonly used method for data analysis: PCA (Pearson, 1901); and secondly, an unsupervised neural network, the Kohonen SOM algorithm (Kohonen, 1995).

2.2.1. Principal components analysis

PCA is used to reduce the dimensionality of data, and to transform interdependent variables into significant and independent components. This statistical method has been extensively described elsewhere (e.g. Gauch, 1982; Legendre and Legendre, 1998). In the present work, fish species abundances were first log ($x + 1$) transformed in order to satisfy the assumptions of PCA and then submitted to centred and normalised PCA, which reduces the influence of species variation (Doledec and Chessel, 1991) and best reveals patterns in data sets

(Gauch, 1982). The analysis was therefore conducted on the correlation matrix. PCA was carried out using version 3 of StatLab (Optima-Deltasoft, 1997).

2.2.2. Kohonen self-organizing map

The Kohonen neural network consists of two layers: the first (input) layer is connected to a vector of the input data set (the 15 fish populations, $n=15$ neurons in the input layer); the second (output) layer forms a map, a rectangular grid with 15 by 10 neurons laid out on a hexagonal lattice ($S=150$ neurons in the output layer) (Fig. 1). The number of cells on the output layer were experimentally defined as the best compromise between representation clarity and computing time. Each neuron of the output layer stores a virtual unit (VU) with species abundance to be computed.

The aim of the SOM algorithm is to visualise in a two dimensional space the SU distribution by way of the VU distribution. VUs which are neighbours on the grid are expected to represent neighbouring clusters of SUs; consequently, distant SUs (according to species abundance) are expected to be distant in the feature space. To achieve this, many distance measures may be

used (e.g. Euclidean distance, Mahalanobis distance, Manhattan distance). The distance measurement method is selected to provide the most accurate data representation on the map. In this study, we used the relative Euclidean distance (RED) (Orloci et al., 1979) in order to equalise the importance of species relative to SUs with high and low total abundances (Ludwig and Reynolds, 1987), aiming to avoid biases due to over or under representation of some species abundances. Then, the distance between two sample units SU_j and SU_k was calculated as follows:

$$RED(SU_j, SU_k) = \sqrt{\sum_{i=1}^S \left[\frac{X_{ij}}{\sum_l X_{lj}} - \frac{X_{ik}}{\sum_l X_{lk}} \right]^2}, \quad (1)$$

with X_{ij} : the abundance of the species i in the SU_j , in other words, the computation was based on the relative proportions of species in the SUs. For this purpose, the input data was standardised relative to total SU abundances during the SOM computation.

The SOM algorithm is an unsupervised learning procedure and can be summarised as follows (see Kohonen, 1995 for more detail):

- The virtual units (VU_k , $1 \leq k \leq S$) are initialised with random samples drawn from the input data set.
- The VUs are updated in an iterative way:
 - A sample unit SU_j is randomly chosen as an input unit.
 - The distance between SU_j and each VU is computed.
 - The virtual unit VU_c closest to the input SU_j , or in other words, the neuron which responds maximally to this input is selected and called ‘best matching unit’ (BMU).
 - The BMU and its neighbours are moved slightly towards the input unit SU_j .

The above described training procedure was broken down into two phases as previously defined by Giraudel et al. (2000):

- Ordering phase (the 2000 first steps): the VUs are highly modified in a wide neighbourhood of the BMU.

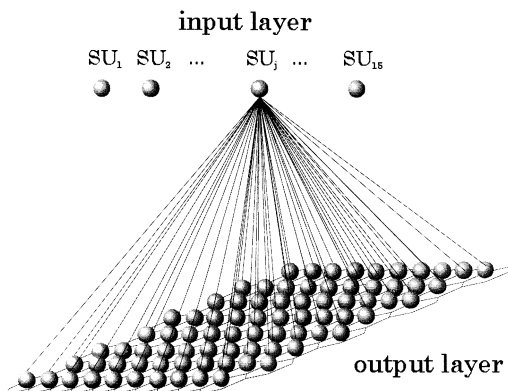


Fig. 1. Representation of the non-supervised artificial neural network (i.e. Kohonen neural network), showing the input neurons and the output neurons organised on a rectangular two dimensional grid. Each input neuron is fed by one fish population and the weight computed by each neuron is represented on the output grid.

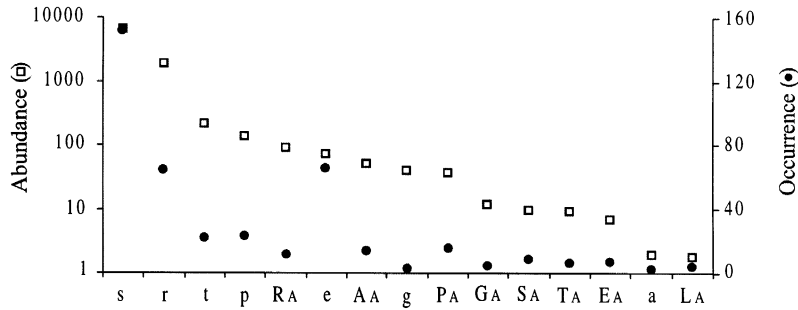


Fig. 2. Occurrence (number of sites where the species is present) and abundance (total number of individuals) of the 15 fish species. Species abundances were represented in log scale to avoid an undue influence of the most abundant species on the figure. Small letters (x) represent 0+ fish populations and capitals followed by 'A' correspond to adults (XA). a, AA: bleak (*Alburnus alburnus*); e, EA: pike (*Esox lucius*); g, GA: gudgeon (*Gobio gobio*); LA: Pumpkinseed (*Lepomis gibbosus*) (only adults were collected during the survey period); p, PA: perch (*Perca fluviatilis*); r, RA: roach (*Rutilus rutilus*); s, SA: rudd (*Scardinius erythrophthalmus*); t, TA: tench (*Tinca tinca*).

- Tuning phase (75 000 steps): only the VUs adjacent to the BMU are lightly modified.

At the end of training, the species abundance are known for each VU, the BMU is determined for each SU, and each SU is set in the corresponding hexagon of the Kohonen map.

The SOM have been computed on a PC with an Intel Pentium PIII-500 using MATLAB software with a program file written by the authors (JLG and SL).

3. Results and discussion

The descriptive analysis of the data matrix considering abundance and occurrence of the 15 species revealed the heterogeneity of the data set (Fig. 2), such a pattern usually being observed in ecological data (Pennington 1996). Within the considered assemblage, two species were numerically dominant (0+ rudd and 0+ roach) and represented respectively 71 and 21% of the 9054 fish collected. The remaining 13 species each accounted for less than 3% of the total fish number. In the same way only three species occurred in more than 9% of the samples with only one widespread species (0+ rudd, present in 21% of the samples) and two moderately occurrent species (0+ roach and 0+ pike, present in 9% of the samples). The remaining species were present

in less than 4% of the samples. Finally, within the entire data matrix (710 records), no fish were found in 409 records (i.e. 58% of the samples). Owing to the different degrees of patchiness and abundance of the species, this data set can be considered as a typical example of complex ecological data (Begon et al., 1996). In the same way, the high data heterogeneity constitutes one of the major limitations which is likely to induce biases for most statistical analysis and representations, as already underlined by Pennington (1996) and ter Braak and Verdonschot (1995).

3.1. Principal component analysis

The PCA allowed the 15 fish species abundances to be taken into account simultaneously aiming to visualise the spatial fish assemblage within the studied area. After logarithmic transformation of the variables, the PCA first and second axis accounted for 10.6 and 10.1% of the total variance respectively (Fig. 3a). This low inertia of the two first axes testified for the complexity of the ecological trends to be visualised and constitutes itself a limitation of the reliability of the PCA results (Gauch, 1982). Considering both correlation and contributions, the first axis showed an opposition between two young fish species (0+ roach and 0+ perch) and two

adult species (adult roach and adult bleak). The second axis showed an opposition between 0+ rudd and the two previous species groups (Fig. 3b). We can also identify an opposition between most adult fish species (bottom right part of Fig. 3b) and 0+ pike. Moreover adults were found independent from 0+ roach and 0+ perch. This opposition or independence between 0+ and adult fishes is ecologically relevant and well known by ichthyologists. In lakes, young fish colonise mainly the shallow littoral areas where they can find both abundant feeding resources and shelter from fish predators (e.g. adult perch) (Werner et al., 1977; Savino and Stein, 1989; Brosse and Lek, 2000). In the same way, the syntopy of 0+ roach and 0+ perch was due to the similar habitats and feeding requirements of the two species (Hammer, 1985; Machacek and Matena, 1997). Moreover, the spatial separation of 0+ rudd and 0+ roach was due to an avoidance of interpopulation competition (Johanson, 1987). Even though these general trends were ecologically relevant,

the consideration of scarce species did not provide useful information: the spatial occupancy of low frequent and low abundant species (e.g. 0+ gudgeon, 0+ bleak, adult rudd) was poorly represented on the PCA $f1 \times f2$ plane. These species were located close to the centre of the plane, without any significant relationship with the two first axes (the other plane representations did not provide better results), whereas it is known that all the species contribute to the community assemblage (Cao et al., 1998). Similarly, irrelevant information was given for 0+ pike, which was found independent from the other young fish populations, whereas actually, this species shares habitat characteristics with 0+ cyprinids (e.g. roach, rudd) and prey on these populations (Eklöv and Hamrin, 1989; Eklöv and Persson, 1996). That drawback could be due to the low abundance of 0+ pike in each sample, indeed cannibalism is common for 0+ pike (Holland and Huston, 1984; Lejolivet and Dauba, 1988), and a given sample never contains more than two individuals.

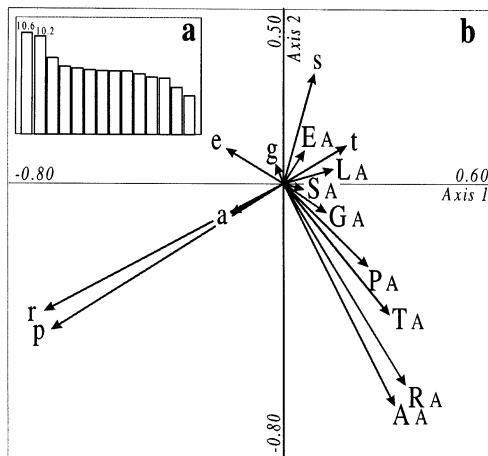


Fig. 3. Results of the normalised principal component analysis (PCA) for the 15 fish species: (a) histogram of eigenvalues; (b) distribution of the 15 species on the $F1 \times F2$ plane. Small letters (x) represent 0+ fish populations and capitals followed by 'A' correspond to adults (XA). a, AA: bleak (*Alburnus alburnus*); e, EA: pike (*Esox lucius*); g, GA: gudgeon (*Gobio gobio*); LA: Pumpkinseed (*Lepomis gibbosus*) (only adults were collected during the survey period); p, PA: perch (*Perca fluviatilis*); r, RA: roach (*Rutilus rutilus*); s, SA: rudd (*Scardinius erythrophthalmus*); t, TA: tench (*Tinca tinca*).

3.2. Kohonen self-organizing map

The results of the distribution of the individuals belonging to the different populations on the Kohonen map are given in Fig. 4. To interpret this map, it should be noted that two neighbouring hexagons contain more closely related individuals than two distant ones. Therefore, the orientation of the map (i.e. the positions of the SUs in one side of the map or in another one) is not important and only the relative positions of the SUs (i.e. the distances between SUs) have to be taken into account. The same general trends as for the PCA were found in the SOM, there being a separation between adults and young fishes in addition to the syntopy of 0+ roach and 0+ perch. However, 0+ pike, was not found independent from the 0+ cyprinids, but logically associated with both 0+ roach and 0+ rudd which constitute usual preys for 0+ pike (Eklöv and Hamrin, 1989; Eklöv and Persson, 1996). A more precise study of this map shows several interesting features, and most of the species exhibited a complex organisation.

tutes a striking example of statistical requirements eclipsing biological common sense. On the contrary, SOM provided a reliable image of the entire assemblage (i.e. considering both dominant and scarce species). Even though SOM is similar to PCA given its ability to reduce the dimensionality of data, its better efficiency to deal with non-linear and heterogeneous data is clearly illustrated. We can therefore consider that SOM, which allows all the species to be considered without biasing the statistical results, constitute a good alternative to common multivariate statistical analysis. Unsupervised network algorithms can therefore be successfully applied to complex ecological data and are able to provide a realistic image of the spatial assemblage of populations without using a priori knowledge about their organisation. Currently, SOM constitute a basic visualisation method for data analysis, and further studies are required to provide more statistical and ecological information from the maps. In this way, the U-matrix methodology (Giraudel et al. submitted), could constitute a fruitful complement to the SOM visualisation ability of complex ecological features.

References

- Aurelle, D., Lek, S., Giraudel, J.L., Berrebi, P., 1999. Microsatellites and artificial neural networks: tools for the discrimination between natural and hatchery brown trout (*Salmo trutta fario*, L.) in Atlantic populations. *Ecol. Model.* 120, 313–324.
- Begon, M., Harper, J.L., Townsend, C.R., 1996. *Ecology: Individuals, Population, and Communities*. Blackwell Science, Oxford.
- Brosse, S., 1999. Habitat, spatial dynamics and fish community structure in lakes, study of lake Pareloup (Aveyron, France). PhD Thesis, University Toulouse, France.
- Brosse, S., Lek, S., 2000. Modelling roach (*Rutilus rutilus*) microhabitat using linear and non-linear techniques. *Freshwater Biol.* 44, 441–452.
- Brosse, S., Dauba, F., Oberdorff, T., Lek, S., 1999a. Influence of some topographical variables on the spatial distribution of lake fish during summer stratification. *Arch. Hydrobiol.* 145, 359–371.
- Brosse, S., Guégan, J.-F., Tourenq, J.-N., Lek, S., 1999b. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. *Ecol. Model.* 120, 299–311.
- Cao, Y., Williams, D.D., Williams, N.E., 1998. How important are rare species in aquatic community ecology and bioassessment. *Limnol. Oceanogr.* 43, 1403–1409.
- Chon, T.-S., Park, Y.S., Moon, K.H., Cha, E., Pa, Y., 1996. Patterning communities by using an artificial neural network. *Ecol. Model.* 90, 69–78.
- Copp, G.H., Gutí, G., Rovny, B., Cerný, J., 1994. Hierarchical analysis of habitat use by 0 + juvenile fish in Hungarian/Slovak flood plain of the Danube river. *Env. Biol. Fish.* 40, 329–348.
- Doledec, S., Chessel, D., 1991. Recent developments in linear ordination methods for environmental sciences. In: Council of Science Research Integration (Ed.), *Trends in ecology*. Research trends Publishers, India. pp. 1–21.
- Eklöv, P., Hamrin, S.F., 1989. Predator efficiency and prey selection: interactions between pike (*Esox lucius*), perch (*Perca fluviatilis*) and rudd (*Scardinius erythrophthalmus*). *Oikos* 56, 149–156.
- Eklöv, P., Persson, L., 1996. The response of prey to the risk of predation: proximate cues for refuging juvenile fish. *Anim. Behav.* 51, 105–115.
- Fore, L.S., Karr, J.R., Wisemann, R., 1996. Assessing invertebrates responses to human activities: evaluating alternative approaches. *J. North Am. Benthol. Soc.* 15, 212–231.
- Gauch, H.G., 1982. *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge.
- Giraudel, J.L., Aurelle, D., Berrebi, P., Lek, S., 2000. Application of the self-organizing mapping and fuzzy clustering to microsatellite data: how to detect genetic structure in brown trout (*Salmo trutta*) populations. In: Lek, S., Guégan, J.F. (Eds.), *Artificial Neuronal Networks, Application to Ecology and Evolution*. Springer-Verlag, pp. 187–202.
- Grossman, G.D., Nickerson, D.M., Freeman, M.C., 1991. Principal component analyses of assemblage structure data: utility of tests based on eigenvalues. *Ecology* 72, 341–347.
- Grossman, G.D., Ratajczak, R.E., Crawford, M., Freeman, M.C., 1998. Assemblage organization in stream fishes: effects of environmental variation and interspecific interactions. *Ecol. Monogr.* 68, 395–420.
- Hammer, C., 1985. Feeding behaviour of roach (*Rutilus rutilus*) and the fry of perch (*Perca fluviatilis*) in Lake Lankau. *Arch. Hydrobiol.* 103, 61–74.
- Holland, L.E., Huston, M.L., 1984. Relationship of Young-of-the-Year northern pike to aquatic vegetation types in backwaters of the upper Mississippi river. *North Am. J. Fish Mgmt.* 4, 514–522.
- Johanson, L., 1987. Experimental evidence for interactive habitat segregation between roach (*Rutilus rutilus*) and rudd (*Scardinius erythrophthalmus*) in a shallow eutrophic lake. *Oecologia* 73, 21–27.
- James, F.C., McCulloch, C.E., 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Ann. Rev. Ecol. Syst.* 21, 129–166.
- Kohonen, T., 1995. *Self-Organizing Maps*. Springer-Verlag, Heidelberg.
- Lamoureux, N., Olivier, J.M., Persat, H., Pouilly, M., Souchon, Y., Statzner, B., 1999. Predicting community characteristics from habitat conditions: fluvial fish and hydraulics. *Freshwater Biol.* 42, 275–299.

- Legendre, P., Legendre, L., 1998. Numerical Ecology. Elsevier.
- Lejollivet, C., Dauba, F., 1988. Growth and feeding behavior of pike (*Esox lucius*, L.) fry reared in cages in Pareloup reservoir. *Annls. Limnol.* 24, 183–192.
- Lek, S., Guégan, J.F., 2000. Artificial neuronal networks, applications to ecology and evolution. Springer-Verlag.
- Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagner, S., 1996. Application of neural networks to modeling nonlinear relationships in ecology. *Ecol. Model.* 90, 39–52.
- Ludwig, J.A., Reynolds, J.F., 1987. Statistical Ecology, a Primer on Methods and Computing. John Wiley & sons.
- Machacek, J., Matena, J., 1997. Diurnal feeding patterns of age-0 perch (*Perca fluviatilis*) and roach (*Rutilus rutilus*) in steep-sided reservoir. *Arch. Hydrobiol.* 49, 59–70.
- Mastrorillo, S., Dauba, F., Belaud, A., 1996. Utilisation des microhabitats par le vairon, le goujon et la loche franche dans trois rivières du sud-ouest de la France. *Annls. Limnol.* 32, 185–195.
- Melssen, W.J., Smits, J.R.M., Rolf, G.H., Kateman, G., 1993. Two-dimensional mapping of IR spectra using a parallel implemented self-organizing feature map. *Chemom. Intell. Lab. Syst.* 18, 195–204.
- Nelva, A., Persat, H., Chessel, D., 1979. Une nouvelle méthode d'étude des peuplements ichtyologiques dans les grands cours d'eau par échantillonnage ponctuel d'abondance. *C. R. Acad. Sci. Paris Serie III* 289, 1295–1298.
- Optima-Deltasoft, 1997. Statlab by SLP. Le logiciel d'exploitation de données.
- Orloci, L., Rao, C.R., Stitler, W.M., 1979. Multivariate analysis in ecological work. Statistical ecology series, USA.
- Pearson, K., 1901. On lines and planes of closest fit to a system of points in space. *Philosoph. Mag.* 2, 557–572.
- Pennington, M., 1996. Estimating the mean and variance from highly skewed marine data. *Fish Bull.* 94, 498–505.
- Persson, L., Greenberg, L.A., 1990. Interspecific and intraspecific size class competition affecting resource use and growth of perch, *Perca fluviatilis*. *Oikos* 59, 97–106.
- Savino, J.F., Stein, R.A., 1989. Behavioural interactions between fish predators and their prey; effects of plant density. *Anim. Behav.* 37, 311–321.
- ter Braak, C.J.F., Verdonschot, F.M., 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquat. Sci.* 57, 254–289.
- Townsend, C.R., Doledec, S., Scarsbrook, M.R., 1997. Species traits in relation to temporal and spatial heterogeneity in streams: a test of habitat templet theory. *Freshwater Biol.* 37, 367–387.
- Werner, E.E., Hall, D.J., Laughlin, D.R., Wagner, D.J., Wilsman, L.A., Funk, F.C., 1977. Habitat partitioning in a freshwater fish community. *J. Fish. Res. Bd. Can.* 34, 360–370.