PATTERNING AND CLUSTERING ECOLOGICAL ASSEMBLAGES

Muriel Gevrey[(1)], Sovan Lek[(1)], Thierry Oberdorff[(2)] & Young Seuk Park[(3)]
(1) LADYBIO, UMR 5172, CNRS-Université Paul Sabatier,
Toulouse, France
gevrey@cict.fr
(2) Institut de Recherche pour le Développement (IRD), MNH, DMPA USM 403,
Paris, France
(3) Department of Biology, Kyung Hee University, Seoul, Korea

Abstract – *The use of advanced modelling methods in ecology expands as ecological data accumulate and increase in complexity. Artificial neural networks and in particular Self-Organizing Map (SOM) have become very popular for analysing particular kinds of ecological datasets. Using the dataset of the distribution of freshwater fish species in France, this paper shows the potential of SOM in ecological modelling and more precisely for patterning French fish assemblages. This paper show also how SOM results can be used by end-users as tool for ecosystems managers by their implementation in a software.*

Key words – Self Organizing Map; Ecosystem Management Software; Community Ecology; France; Fish

1 Introduction

In ecology, the amount and complexity of data increase. Powerful methods are then required for their analysis. For many years, classical statistical methods have been used to analyse ecological data, for example, the multiple linear regression [1], the canonical correspondence analysis [2] or the multiple dimensional scaling [3].
Alternative methods have then been developed which are better adapted to complex data. They are increasingly being applied in ecological research, for example, the genetic algorithm [4], the classification and regression trees [5] or the artificial neural network (ANN) [6, 7, 8].
The multi-layered feed-forward neural network, trained with a back propagation algorithm [9] is the most common artificial neural network method used in ecology [10, 11, 12]. However, the Self-Organizing Map (SOM) [13] method becomes very popular for analysing particular kinds of ecological dataset. The SOM is an efficient method for analysing systems ruled by complex non-linear relationships and provides an alternative to traditional statistical methods to classify ecological data [14]. SOM have been used successfully in ecology for instance, for patterning communities [15], for the assessment of water quality [16], for the classification of rainfall variability [17] or for the modelling of population dynamics of aquatic insects [18].
The SOM is often used to project the dataset in a non-linear way onto a topological rectangular grid arranged as a hexagonal lattice that is called a map. In addition to this visualisation, there is also an underlying SOM model whose outputs can de displayed in several different ways to reveals different types of information.
In this paper, the potential of SOM in ecological modelling is illustrated with a dataset of the freshwater fish species distribution in France. The aim of this study was to patterning fish species distribution in French rivers and to evaluate the relative importance of several environmental

factors in influencing organization and structure of fish assemblages. Moreover, the results was used in a software suggesting a set of tool for water management and water policies in order to facilitate the assessment of ecological quality and perturbations of stream ecosystems. This software is usable by end-users as scientists and ecosystem managers.

2 Material and Methods

2.1 Data

The data, extracted from the database held by the Conseil Supérieur de la Pêche (Banque Hydrobiologique et Piscicole), were previously analysed by [19, 20]. The dataset is constituted of 668 reference sites (Fig.1) collecting during a period of 13 years of survey (1985-98). The selection of the reference sites was carried out by regional experts (fish biologists) on the basis of water quality map inspection and field reconnaissance. In the dataset, 40 species were identified (Table 1).

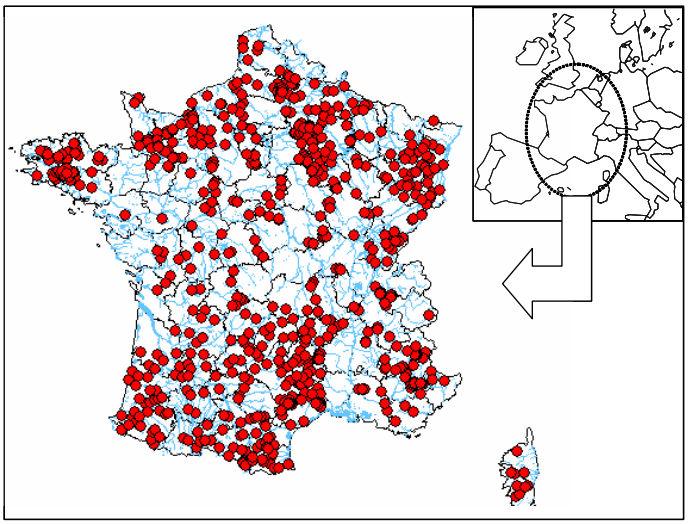

Figure 1 Map of France showing all sampling sites.

Eight abiotic environmental variables were also measured at each site: slope (%), elevation (m), July mean daily maximum air temperature (°C; JulTemp), January mean daily maximum air temperature (°C; JanTemp), stream width (m), mean depth (m), distance from headwater source (km), and catchment area of the basin (km^2). The slope and elevation were derived from topographic maps, and the distance from the source and the catchment area were measured using a digital palimeter on a 1:1,000,000-scale map. A detailed description of all these environmental variables is given in [19]. These variables are known to be the most consistent in structuring fish assemblages under natural conditions.

Abbreviations	Common name	Scientific name
ABB	Bream	*Abramis brama*
ALB	Schneider	*Alburnoides bipunctatus*
ALA	Bleak	*Alburnus alburnus*
ANA	European eel	*Anguilla anguilla*
BAB	Barbel	*Barbus barbus*
BAM	Mediterranean barbell	*Barbus meridionalis*
BLF	Freshwater blenny	*Blennius fluviatilis*
BLB	Silver bream	*Blicca bjoerkna*
CAA	Goldfish	*Carassius auratus*
CAC	Crucian carp	*Carassius carassius*
CHN	Common nase	*Chondrostoma nasus*
CHT	Soiffe	*Chondrostoma toxostoma*
COG	Bullhead	*Cottus gobio*
CYC	Common carp	*Cyprinus carpio*
ESL	European pike	*Esox lucius*
GAF	Mosquitofish	*Gambusia affinis*
GAC	Threespined stickleback	*Gasterosteus aculeatus*
GOG	Gudgeon	*Gobio gobio*
GYC	Ruffe	*Gymnocephalus cernua*
ICM	Black bullhead	*Ictalurus melas*
LAP	Brook lamprey	*Lampetra planeri*
LEG	Pumpkinseed	*Lepomis gibbosus*
LED	Belica	*Leucaspius delineatus*
LEC	Chub	*Leuciscus cephalus*
LEL	Dace	*Leuciscus leuciscus*
LES	Varione	*Leuciscus souffia*
LOL	Burbot	*Lota lota*
MIS	Largemouth bass	*Micropterus salmoides*
NEB	Stone loach	*Nemacheilus barbatulus*
PEF	Perch	*Perca fluviatilis*
PHP	Minnow	*Phoxinus phoxinus*
PUP	Ninespined stickleback	*Pungitius pungitius*
RHS	Bitterling	*Rhodeus sericeus*
RUR	Roach	*Rutilus rutilus*
SAS	Atlantic salmon	*Salmo salar*
SAT	Brown trout	*Salmo trutta fario*
SCE	Rudd	*Scardinius erythrophthalmus*
STL	Zander	*Stizostedion lucioperca*
THT	Grayling	*Thymallus thymallus*
TIT	Tench	*Tinca tinca*

Table 1. List of 40 species identified in dataset.

To find the biogeographical distribution patterns of fish species in French rivers, the dataset was introduced to the SOM. The densities of species were scaled between 0 and 1 in the range of the minimum and maximum values within abundance of species, after a log-transformation process in order to reduce variations in densities.

2.2. SOM for ecological modelling

Ecological data are commonly constituted of a set of objects described by several descriptors. The object, as in our study, is sample sites and the descriptors is a set of species found in each site (abundance of species) or also a set of environmental variables characterizing the sites. One of the goals achieved by the application of SOM in ecology is the classification of the sample sites (the objects) in the SOM map (the output layer of the network) according to the similarities between the species (descriptors).
The SOM is usually constituted of two layer (the input and the output), linked by weights associated to connection intensities. The output layer is also called the map. It is a two-dimensional network of neurons arranged on a hexagonal lattice. This map consists of N neurons ($35=7\times5$ in this study) which usually constitute a 2D grid for better visualization. Before the learning process, weight vectors constituted of as much unit as input neuron (so as species in the data) are assigned randomly to each neuron of the output layer. The output neurons are considered as virtual units to represent typical patterns of the input dataset assigned to their units after the learning process. The training starts when an input vector is sent through the network. This input vector is one of the sample sites. Each neuron of the output layer computes the summed distance between the weight vectors and the input vector. Among all the output neurons, the best matching unit (BMU) which has the minimum distance between weight and input vectors becomes the winner. The weight vectors of the BMU and its neighbourhood units are then updated by the SOM learning rule. The sample site is assigned to the BMU. All the samples sites are then classified in the unit of the SOM map at the end of the training.
A cluster analysis can be applied to the trained SOM to define several levels of groups of virtual units. In this study a hierarchical cluster analysis was used with Ward's linkage method.
The contribution of each species in the classification of the sites and in the cluster structures helps to define fish community assemblages. The value of each input variable calculated during the training process is displayed in each neuron on the trained SOM map on a grey scale.
Finally, using the mean value of each environmental variable in each virtual unit of the SOM map following [21], it is possible to analyse the relationships between biological and environmental variables. These mean values assigned on the SOM map are visualised with a grey scale, and then are compared with maps of sampling sites as well as species maps.

3 Results and Discussion

3.1 Assemblage patterning

The classification of our sampling sites and the patterning of the fish assemblages on the national scale were realised by training the SOM (Fig. 2a). Using a hierarchical clustering analysis with the Ward method after the learning process of the SOM, several clusters were found (Fig. 2b). The numbers in the dendrogram in Fig. 2b corresponds to the number of the units of the SOM map. The weight vector of each unit represents a typical assemblage composition of samples. Two major clusters (I, II) were divided into two subclusters (IA and IB, IIA and IIB). Six clusters appeared finally at the distance level of 0.8 (IA-IIBb).

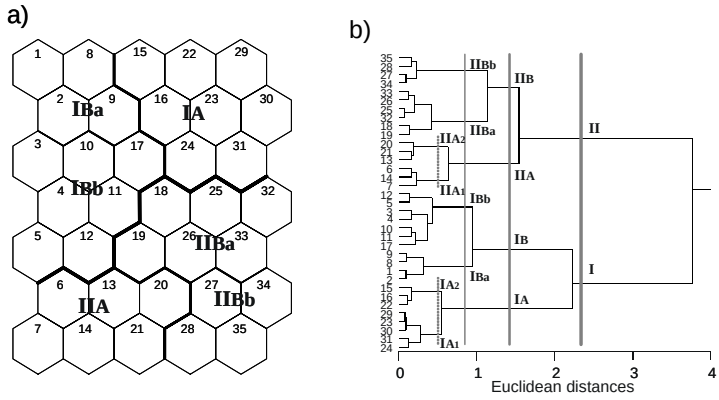

Figure 2 Classification of fish assemblages on the SOM map (a) and hierarchical classification of SOM units using Ward's algorithm (b). Each unit of the map represents a typical assemblage composition of samples by taking its weight vector.

Figure 3 displays the distribution of each fish species in each neuron of the SOM map in a grey scale, by visualizing the weight vectors of the SOM. For convenience of interpretation, the values of weights were rescaled between 0 (lighter colour) and 1 (darker colour). In Fig. 3, different map are obtained according to the species studied showing different distribution patterns. For instance, the species *Salmo trutta fario* (SAT) and *Thymallus thymallus* (THT) were the most abundant in the upper right areas of the map (cluster IA). *Cottus gobio* (COG), *Lampetra planeri* (LAP) and *Salmo salar* (SAS) are in the upper left areas (cluster IBa). *Nemacheilus barbatulus* (NEB), *Phoxinus phoxinus* (PHP) and *Pungitius pungitius* (PUP) were in the left areas (cluster IBb). 10 species including *Alburnoides bipunctatus* (ALB), *Esox lucius* (ESL), and *Leucaspius delineatus* (LED) were in the lower left areas (cluster IIA). *Barbus meridionalis* (BAM), *Blennius fluviatilis* (BLF), *Leuciscus souffia* (LES), and *Lota lota* (LOL) were in the middle right areas (cluster IIBa) and 16 species including *Blicca bjoerkna* (BLB) and *Perca fluviatilis* (PEF) were in the lower right areas (cluster IIBb). Based on these distribution maps, we could find the species distribution patterns at different sampling sites.

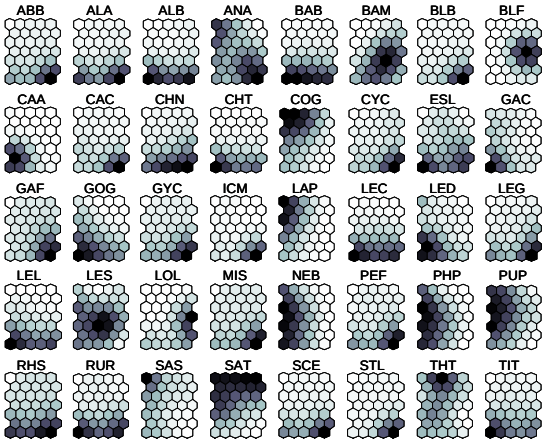

Figure 3 Visualization of relative abundance of species calculated in the trained SOM in grey scale. The values were calculated during the learning process. Dark represents high values of abundance, whereas light is for low values. The acronyms of the species are presented in Table 1.

To understand the relationships of the environmental variables with the fish assemblages but also with the classification of sampling sites in the SOM, mean values of each environmental variable were calculated and visualized in the trained SOM map (Fig. 4). High values of each variable are represented by dark colours, whereas light colours are used for low values. Environmental variables showed a clear gradient distribution on the SOM map. The catchment area, the distance from the source, the width and the depth of the sampling areas were the highest values in the lower right areas of the SOM map (cluster IIBb), whereas lower values appeared in the upper areas (cluster IA). In contrast, the slope and the altitude were the highest in upper left area of the SOM map (cluster IA), while lower values were in the lower right areas (cluster IIBb). Meanwhile, temperatures in January and July were higher in the middle right areas of the map, although the distribution gradients were not clear.

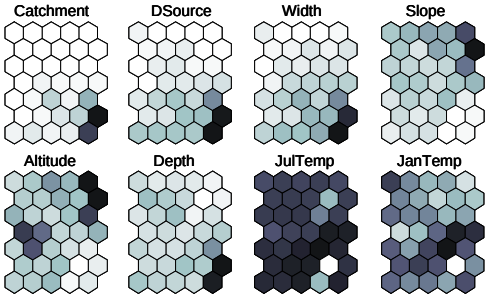

Figure 4 Visualization of environmental variables on the SOM map trained with fish assemblages. The mean value of each variable was calculated in each output unit of the trained SOM. Dark represents a high value, while light is low. Catchment; catchment area, DSource; distance from source, JulTemp; maximum temperature in July, and JanTemp; maximum temperature in January.

Overall the SOM showed six clusters of fish assemblages, highly related to longitudinal river gradient. These characteristics support the fish zonation theory in European continental river [22]. Recently [19] developed a probabilistic model characterizing fish assemblages of French rivers with environmental variables. They showed that the probability of occurrence is highly dependent on the longitudinal gradient. Our findings on the occurrence patterns for most species agree with their results, although there are some small differences for a few species. The results showed also a significant relationship between the trophic structure of the assemblages and river size such as catchment area, width, depth, and distance from head water source, supporting the river continuum concept [23].

3.2 End-user tool

These results have been directly used in a software created for an European project, PAEQANN ("Predicting Aquatic Ecosystem Quality using Artificial Neural Network"; N° EVK1-CT1999-00026) which had for goal to develop general methodologies, based on advanced modelling techniques, for predicting structure and diversity of key aquatic communities under natural and man-made disturbances. The software had the objective to suggest a set of tools for water management and water policies in order to facilitate the assessment of ecological quality and perturbations of streams ecosystems. This tool allow also scientists and ecosystem managers to consult the occurrence patterns of organisms in stream based on the database used in the tool, visualise the results of patterning and predicting models with existing data, and provide the possibility of testing the new data based on models developed with existing data.
SOM was used in this tool as an ordination method to summarize the variability of the data. Thus sampling sites could be arranged on the reduced dimensions, so that these arrangements optically summarize the spatial variability of their biological and environmental features.

Fig. 5 is the result frame displayed when the user has selected the ordination button for the French fish communities. The correspondence between the sampling sites represented on the geographic map and the cluster they belong to, according to the colour on the SOM map, can be seen. Several level of cluster can be observed. Moreover, the environmental variables can also be visualized on the SOM map.

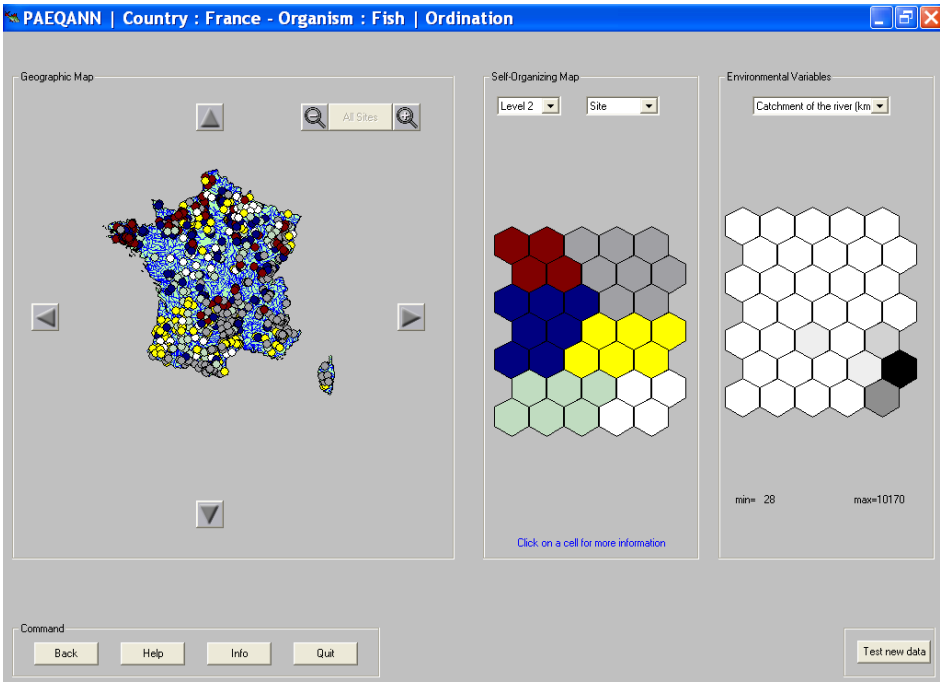

Figure 5 Example of an ordination window indicating fish communities in France

Finally, the user can test new community data with trained SOM by presenting corresponding values. The tested results are then indicated on the corresponding unit of the SOM map with a black circle. This tool can be downloaded at http://aquaeco.ups-tlse.fr.

4 Conclusion

The national fish distribution characteristics were efficiently visualized on reduced dimensions through the SOM. The results confirm major concepts in fish ecology such as the river continuum concept. SOM seem to be a powerful analytical tool for identifying habitat and species grouping. Moreover, the implementation of the SOM result in a tool for ecosystem manager is very encouraging for the future use of SOM in ecology.

References

[1] N.A. Binns and J.P. Eiserman (1979), Quantification of fluvial trout habitat in Wyoming. *Transactions of the American Fisheries Society*, **198**: p. 215-228.

[2] C. ter Braak (1987), The analysis of vegetation-environment relationships by canonical correspodence analysis. *Vegetatio*, **69**, 69-77.

[3] E. Guilherme and R. Cintra, (2001): Effects of intensity and age of selective logging and tree girdling on an understory bird community composition in Central Amazonia, Brazil. *Ecotropica*,**7**, 77-92.

[4] P. Chaves, T. Kojiri and Y. Yamashiki (2003) , Optimization of storage reservoir considering water quantity and quality. *Hydrological Processes,* **17**, 2769-2793.

[5] J. Gregor, N. Garrett, B. Gilpin, C. Randall and D. Saunders (2002), Use of classification and regression tree (CART) analysis with chemical faecal indicators to determine sources of contamination. *New Zealand Journal of Marine and Freshwater Research,* **36.**, 387-398.

[6] H. R. Maier and G. C. Dandy (2000), Neural networks for the prediction and forecasting of water resource variables: A review of modelling issues and applications. *Environmental Modelling & Software* **15**, 101-124.

[7] S. C. Michaelides, C. S. Pattichis, and G. Kleovoulou (2001), Classification of rainfall variability by using artificial neural networks. *International Journal of Climatology,* **21**, 1401-1414.

[8] R. Cereghino, J. L. Giraudel, and A. Compin (2001), Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self organizing maps. *Ecological Modelling,* **146**, 167-180.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams (1986), Learning internal representations by error propagation, pp. 318-362. In D. E. J. M. Rumelhart (Ed.): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, Massachussets, USA.

[10] T. A. Clair, and J. M. Ehrman (1998), Using neural networks to assess the influence of changing seasonal climates in modifying discharge, dissolved organic carbon, and nitrogen export in eastern Canadian rivers. *Water resources research,* **34**, 447-455.

[11] O. Antonic, J. Krizan, A. Marki and D. Bukovec (2001), Spatio-temporal interpolation of climatic variables over large region of complex terrain using artificial neural networks. *Ecological Modelling,* **138**, 255-263.

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

[13] T. Kohonen (2001), *Self-Organizing Maps*. Springer-Verlag. Heidelberg.

[14] S. Lek and J.-F. Guegan (2000), *Artificial Neuronal Networks, Application to Ecology and Evolution.* Springer-Verlag. Heidelberg.

[15] T. S. Chon, Y. S. Park, K. H. Moon and E. Y. Cha (1996), Patternizing communities by using an artificial neural network, *Ecological Modelling*, **90**, 69-78.

[16] M. Gevrey, F. Rimet, Y. S. Park, J. L. Giraudel, L. Ector, and S. Lek (2004), Water quality assessment using diatom assemblages and advanced modelling techniques. *Freshwater Biology* **49**, 208-220.

[17] S. C. Michaelides, C. S. Pattichis and G. Kleovoulou (2001), Classification of rainfall variability by using artificial neural networks. *International Journal of Climatology* **21**, 1401-1414.

[18] M. Obach, R. Wagner, H. Werner and H-H. Schmidt (2001), Modelling population dynamics of aquatic insects with artificial neural networks. *Ecological Modelling* **146**, 207-217.

[19] T. Oberdorff, D. Pont, B. Hugueny and D. Chessel (2001), A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. *Freshwater Biology,* **46**, 399-415.

[23] T. Oberdorff, D. Pont, B. Hugueny and J.P. Porcher (2002), Development and validation of a fish-based index for the assessment of rivers "health" in France. *Freshwater Biology,* **47**, 1720-1735.

[20] Y.S. Park, R. Cereghino, A. Compin and S. Lek (2003), Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling,* **160**, 265-280.

[21] M. Huet (1954), Aperçu des relations entre la pente et les populations piscicoles des eaux courantes. *Schweiz Z Hydrol,* **11**, 331-351.

[22] R. Vannote, G.W. Minshall, K.W. Cummins, J.R. Sedell and C.S. Cushing (1980), The river continuum concept. *Canadian Journal of fisheries and Aquatic Sciences,* **37**, 130-137.