&ECOHYDROLOGY
HYDROBIOLOGY

Vol. 4
No 2, 139-146
2004

Patterns of fish assemblages in tropical streamlets using SOM algorithm and conventional statistical methods

Tadeusz Penczak[1], Sovan Lek[2], Francisco Godinho[3], Angelo Antonio Agostinho[4]

[1]Department of Ecology and Vertebrate Zoology, University of Łódź,
90-237 Łódź, 12/16 Banacha Str., Poland,
e-mail: penczakt@biol.uni.lodz.pl
[2]LADYBIO, UMR CNRS - Université Paul Sabatier,
118 Route de Narbonne, 31062 Toulouse cedex, France.
e-mail: lek@cict.fr
[3]National Water Council,
Rua de S. Domingos a Lapa, n. 26, 1200-835 Lisboa, Portugal.
e-mail: francisco.godinho@sg.mcota.gov.pl
[4]NUPELIA, Universidade Estadual de Maringa, Maringa, PR, Brazil.
e-mail: agostinhoaa@nupelia.br

Abstract

Six sites located on the Caracu River and five sites located on the Água do Rancho River (small tributaries of the Paraná River, Brazil) were chosen for quantitative electrofishing sampling to examine patterns in fish assemblage structure using the dualism ordination technique developed by Romaniszyn. Canonical correspondence analysis was also used for verification of the data. A similar separation of sites from both streams by both methods confirmed the Romaniszyn method (RD) was useful for the study of ordination. The RD was also tested using Self-Organizing Maps (SOM), which are a variant of the Artificial Neural Network (ANN) methods. Similarly, the SOM confirmed that sites belonging to the two investigated streams were distinct, but also unveiled a minor weakness of the RD resulting from its linear character. However, in spite of this weakness, we recommend the RD for assemblage analysis by scientists unfamiliar with canonical correspondence and ANN analyses.
Key words: Paraná River, tropical fish population, assemblage structure, Self-Organizing Map, canonical correspondence analysis, Romaniszyn diagram.

1. Introduction

Investigations to develop new and improved techniques of analysing assemblage structures are still in progress. According to Matthews (1998), it is desirable to examine the results from several techniques to determine if fish assemblages occur in distinct groups or randomly in multivariate space. Comparative studies are useful, but this dataset is not experimental in that it has known properties. Three approaches discussed in this paper may have similar weaknesses and misrepresentations of data. However, when similar results are obtained with the different methods, the credibility and validity of all the methods are increased.

A hypothesis regarding differences in fish assemblages in two Paraná streamlets has already been analysed by the Romaniszyn method (Penczak *et al.* 1994) and canonical correspondence analysis (Penczak *et al.* 2002). In addition, the Romaniszyn method has been subsequently verified by the Kohonen Self-Organizing Map (SOM) algorithm (Kohonen 1982, 2001). Several authors (e.g., Chon *et al.* 1996, 2000; Giraudel *et al.* 2000; Giraudel, Lek 2001; Brosse *et al.* 2001; Park *et al.* 2001) have acknowledged this latter method as more objective.

Romaniszyn (1970) described his ordering method when gradient analyses were rarely applied in ecology, but the paper was unfortunately published in Polish and overlooked by later investigators. Attempts at reviving and developing the method have received a number of rejections, though some papers using the method eventually appeared in international journals (e.g., Penczak *et al.* 1994, 2000, 2002; Adamczyk et al. 2004). The verification of the Romaniszyn method (RD) with the canonical correspondence analysis (CCA) has proved promising for the comparison of fish assemblage structure in the two Paraná River streamlets (Penczak *et al.* 2002). CCA and RD portrayed very similar distributions of sites and species, though a limitation of the latter method was that it did not explain the variation in assemblage composition. Hence, the authors concluded that more complex gradients in assemblage structure might limit the application of RD. The same data set (Penczak *et al.* 1994) was analysed in this study by the Self-Organizing Map (SOM), which has been used in a few studies to depict relationships within ecological assemblages (Chon *et al.* 1996, 2000; Giraudel *et al.* 2000; Giraudel, Lek 2001; Brosse *et al.* 2001; Park *et al.* 2001).

The goal of this study was is to demonstrate the use of three different methods - SOM, CCA, and RD - to analyse complex, nonlinear fish assemblage data sets. We tested whether groups of neurons distinguished by SOM on the basis of fish abundance could confirm or even improve groupings of sites that differed in qualitative and quantitative species composition.

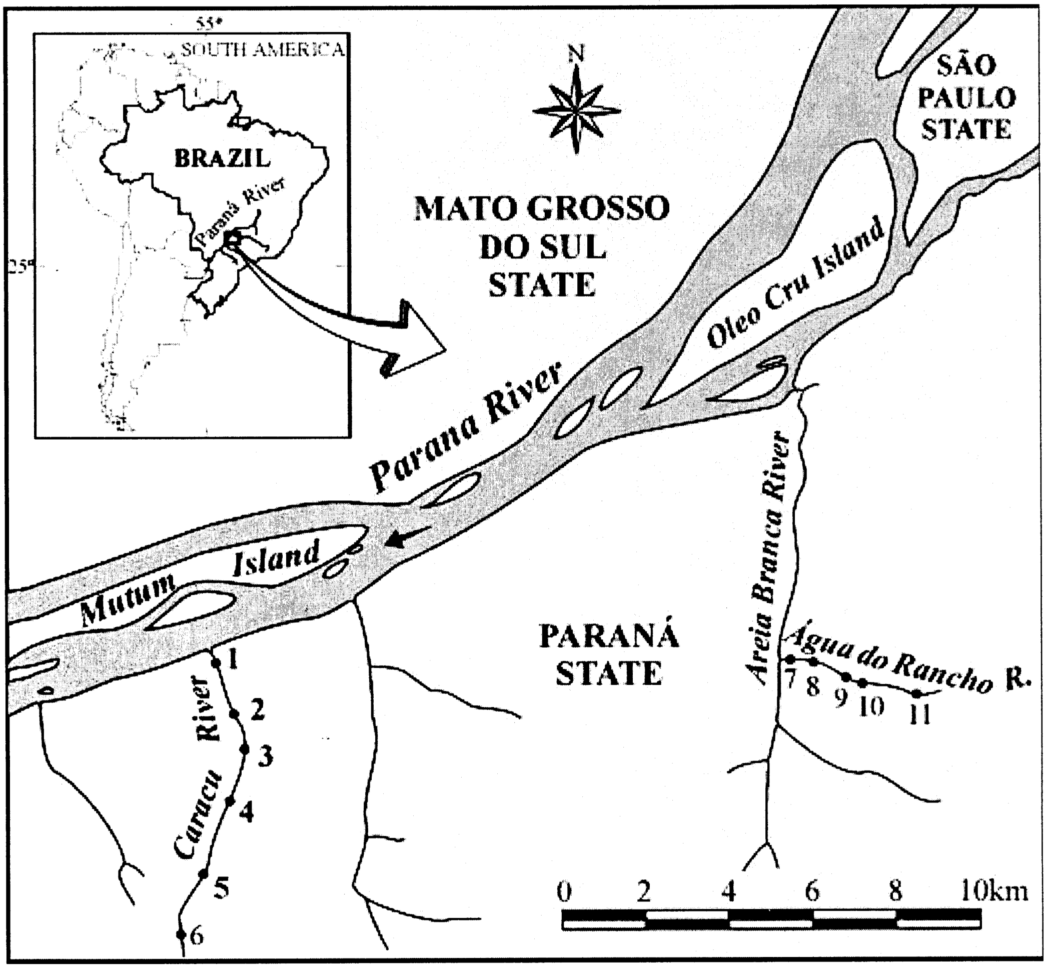

Fig. 1. Study area with sites indicated.

2. Study Area, Materials and methods

Six sites were located on the Caracu River, and five on the Água do Rancho River, both of which are small tributaries of the Paraná River, Brazil (Fig. 1). The Caracu River (6.8 km long) empties directly to the Parana River, whereas the Água do Rancho (4 km long) empties to the Areia Branca River at 7.8 km from its outlet to the Parana River. Adjacent landscapes, soil, site morphology, and some water chemistry parameters from these rivers are described elsewhere (Penczak *et al.* 1994; Agostinho, Penczak 1995).

Assemblages consisted of 1260 fish, representing 27 taxa from 14 families (Penczak *et al.* 1994). For studies with the three ordering methods, the 20 most abundant species were always selected (see Tables II, III in Penczak *et al.* 1994 and Table I in Penczak *et al.* 2002). Rare taxa, which occurred in less than three sites or had total abundances of less than three specimens, were removed to prevent analysis distortion. Neither abundance nor environmental variable data were transformed. Fish sampling was conducted by three successive electrofishings (three-pass depletion, wading with two anode-dipnets) at each site in October 1992. Sample reaches were blocked with nets at upper and lower ends during sampling; each successive pass was made with a constant unit of effort. Because of low conductivity in the Agua do Rancho Stream, the water was continuously salted during electrofishing. A description of this procedure is presented in Penczak et al. (1997, 2003). The RD method is based on the principle of dualism, which consists of ordering the data twice using the similarity definition (*s*):

$$s = w/a + b - w * 100,$$

where w is the total of the lower number of specimens of each pair of species common for two given analysis collections, a is the total of specimens of a species at the site (or the total number of sites with species A), and b is the total number of specimens of a species in a site (or the total number of sites with species B).

First, similarity was estimated between each pair of sites ('columns' in an initial table with sites and species collected), and second between each pair of species ('rows' in the initial table). Similarities were obtained between columns in one table and then rows in another table followed by construction of two, branched, two-dimensional dendrites. They were then converted to linear dendrites (i.e. ordination axes) respecting the principle that the weakest connection must be 'broken' to insert sites/species that were not in linear order. This procedure and how to distinguish between clusters of sites and/or species is detailed in Penczak *et al.* (1994, 2000). In accordance with Romaniszyn's method (Romaniszyn 1970), all sites and species that were not included in clusters were treated separately during analysis.

The CCA was chosen as a gradient analysis because two matrices were available in this research, i.e. one with species abundance and another with environmental variables (Agostinho, Penczak 1995). The algorithm used in the CCA is reciprocal weighted-averaging of the number of species x sites matrix, together with linear least-squares regression on the environmental variables (ter Braak 1987; Palmer 2000). First, a general CCA was performed with all the environmental variables. Second, an additional CCA was performed with the environmental variables selected by a forward selection procedure available in CANOCO (version> 3.1) as recommended by ter Braak, Verdonschot (1995). For both CCAs, a Monte Carlo simulation with 1000 permutations was used to test the significance of the fish-environment relationships (ter Braak 1987, 1990). Additional details on the application of CCA to verify the RD can be found in Penczak *et al.* (2002).

The Self-Organizing Map of Kohonen (SOM) belongs to the Artificial Neural Network (ANN) class of techniques and is one of the best known neural networks with unsupervised learning rules. It is described in about 4000 research articles and books (Kohonen 2001) and is being increasingly used by ecologists (e.g. Chon *et al.* 1996; Giraudel *et al.* 2000; Giraudel, Lek 2001; Brosse *et al.* 2001). The created map has a regular grid of processing units (neurons). The procedure produces a model of multidimensional observations with optimal accuracy, so that similar items are close to each other and dissimilar items far apart. Fitting of the model vectors is done by a sequential regression process, where $t = 1,2, \ldots$ is the step index: for each sample $\boldsymbol{x}(t)$, first the winner index c (best match) is identified by the condition:

$$\forall i,\ \| \mathbf{x}(t) - \mathbf{m}_c(t) \| \leq \mathbf{x}(t) - \mathbf{m}_i(t) \|.$$

Next, all model vectors or a subset of vectors belonging to nodes centred around node $c = c(x)$ are updated as:

$$\mathbf{m}_i\,(t + 1) = \mathbf{m}_i(t) + h_{c(x),i}\,(\mathbf{x}(t) - \mathbf{m}_i(t))$$

Here, $h_{c(x),i}$ is the "neighbourhood function", a decreasing function of the distance between the i^{th} and c_{th} nodes on the map grid. This regression is usually reiterated over the available samples (Kohonen 2001).

The special importance of SOM for assemblage research is that it allows projection of data in a regular, two-dimensional space, i.e., similarly to the gradient analysis. The SOM consists of two layers:

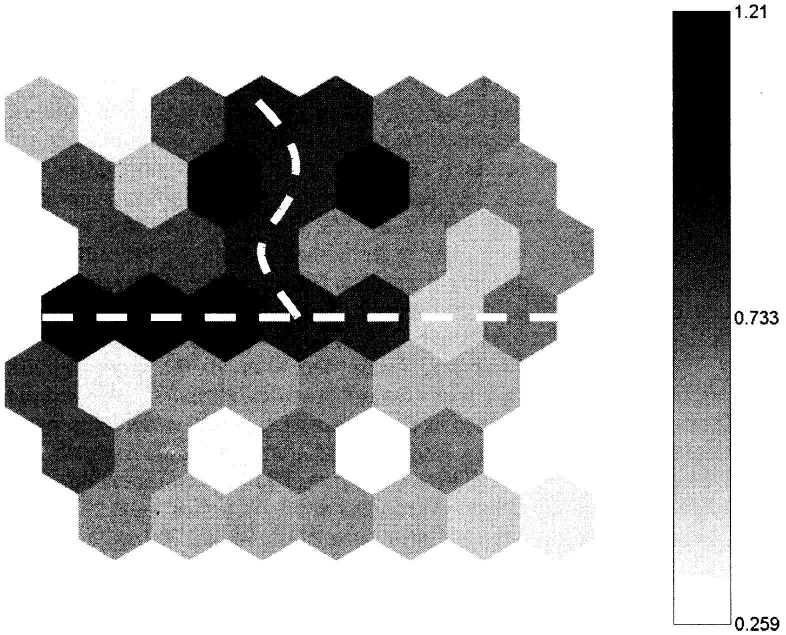

Fig. 2. A self-organizing map formed by 16 hexagons representing virtual units *VUk*. Groups of neurons were distinguished on the basis of the U-matrix. The shading intensity indicates the degree of activation. According to the intensity seen on the U-matrix map, we can distinguish three clusters.

1) the input, which is connected to each vector of the data set, and
2) the output, which forms a two-dimensional array of nodes with weight vectors (Kohonen 1982, 1995, 2001).

For learning, only input units are used. The input layer receives input values from the data matrix, whereas the output layer consists of several output neurons that are usually arranged into a two-dimensional grid for better data visualisation. In this study, the output layer of the SOM consisted of 16 neurons (virtual units *VUk*) arranged into a 4 x 4 hexagonal lattice. During the learning process of the SOM, weights were modified to minimise the distance between weight and input vectors. The learning process was usually done in two phases: first, a rough ordering using a large neighbourhood radius, and secondly, fine-tuning using a small radius. The detailed algorithm of the SOM for ecological applications can be found in Chon *et al.* (1996), Giraudel *et al.* (2000) and Park *et al.* (2001, 2003). The map obtained after learning represented all the observations (*SUs*) assigned to virtual units (*Vuk*), so that similar units were located close to each other and far from those dissimilar. Information relating to species composition was also available for each neuron.

3. Results

First, a U-matrix was used to measure a distance matrix between units of SOM (Fig. 2). The highest values on the right scale indicated the greatest distances between virtual units. In the figure, the dark areas represent large differences between units whereas light areas represent small differences. On the basis of the picture of the U-matrix, three clusters with sites on the SOM map were identified and separated (Fig. 3); note that each two adjoining hexagons in Fig. 3 are separated by one hexagon in the U-matrix (Fig. 2) which indicates degree of similarity between the two. At the end of the learning process, eight hexagons with their site compositions were established among 16 virtual units. Sites sampled in the Caracu River (sites 1-6) were clearly separated from those belonging to the Água do Rancho River (sites 7-11); the former are distributed in virtual units *VU A* and *B* only (one cluster at the bottom of Fig. 3), the latter in *VU C* and *D* (two clusters at the top of Fig. 3). There was no overlap between sites.

The greatest differences were expected between virtual units *VU A1-3* and *VU D1-3*, as documented in the U-matrix (Fig. 2). The sites of the former set of units are distinguished by *Macrolepidogaster* sp., *Astyanax scabripinis*, *Bryconamericus stramineus, Cetopsorhamdia iheringi*, *Phenacorthandia* sp. and *Eigenmania trilineata*, which are dominants or subdominants in the Água do Rancho River, but totally absent in the Caracu River. The sites of the latter set of units have been settled exclusively by *Astynax bimaculatus* (dominant) and five rare species, which are absent in the Água do Rancho River. Site 11(*VU B4* - Água do Rancho) and site 8 (*VU C4* - Caracu) display the greatest similarity, having four common species. Site 5 (*VU C3*) is also quite similar, with three species in common with the sites 8 and 11 (Fig. 3).

At the end of the learning process, 16 virtual units with fish species composition were obtained with twenty taxa displayed on the maps

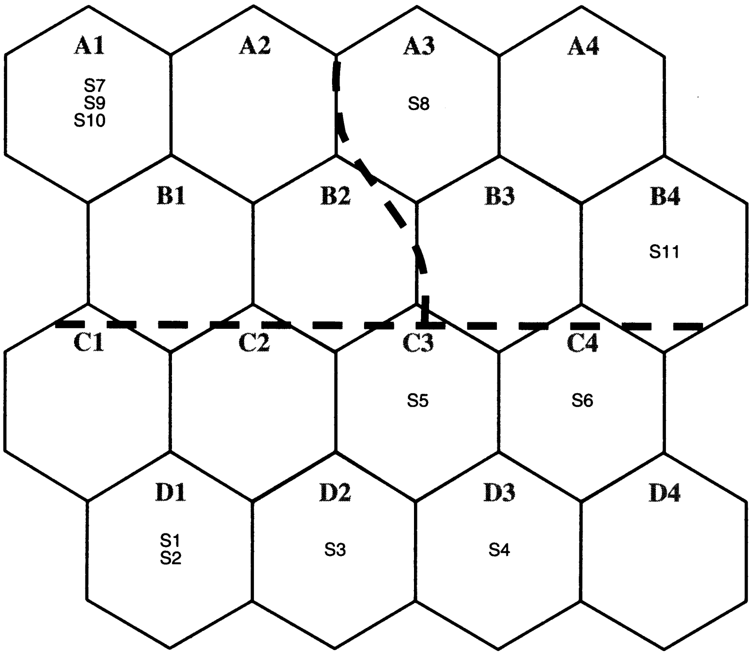

Fig. 3. The Caracu (sites 1 - 6) and Água do Rancho (sites 7 - 11) river sites mapped on the Self-Organizing Map (Fig. 2). Água do Rancho is more heterogeneous (two clusters) than the Caracu (one cluster).

(Fig. 4). All species marked in black on *VU A* and *B* are dominants in the Água do Rancho River, and those marked in black on *VU C* and *D* occur in high numbers only in the Caracu River (Fig. 4). *Gymnotus carapo* and *Hypostomus ancistroides* are the only two taxa that occur at all sites of both streamlets; the former one in moderate abundance in both streams with the latter always in high abundance in the Água do Rancho River and low in the Caracu River. SOMs of the species compositions are a very useful tool for studying fish assemblages in the investigated streams (Fig. 4). From Fig. 4, species that co-occur in the investigated streams can be readily determined, as well as if they are dominant or rare in a given assemblage.

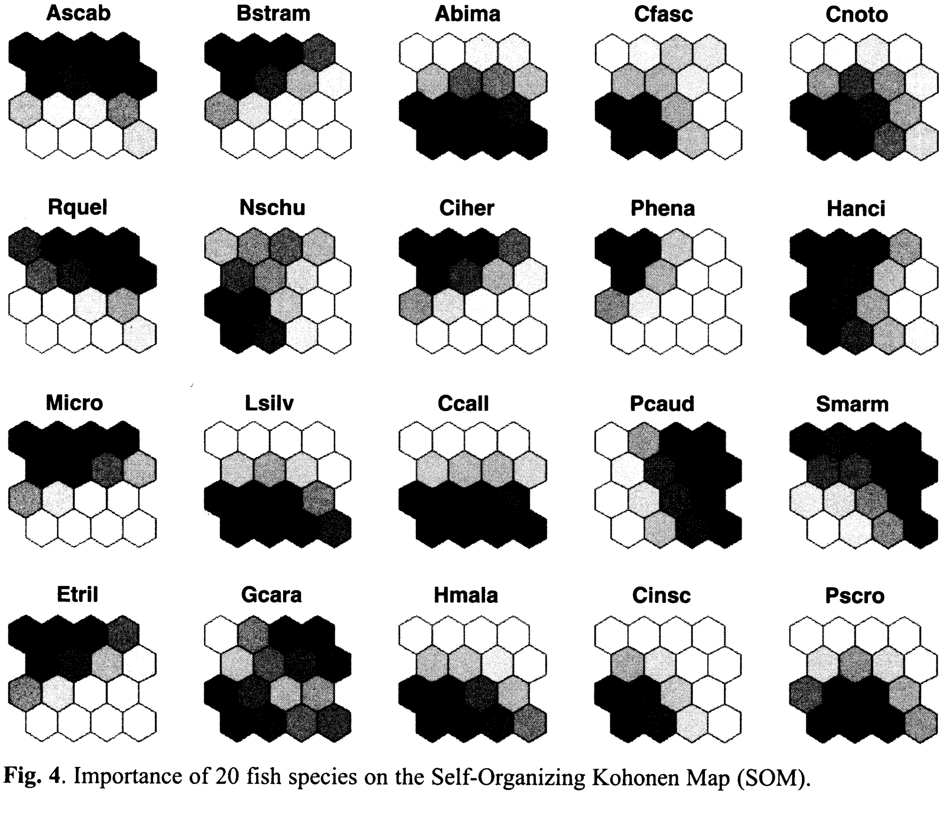

Fig. 4. Importance of 20 fish species on the Self-Organizing Kohonen Map (SOM).

4. Discussion

Artificial Neural Networks (ANN) realistically model functional properties of the human brain. The SOM algorithm used for unsupervised learning is the most popular two-layer type, allowing for reduction of a higher-dimensional space into fewer dimensions as well as improved visualisation and interpretation of the data structure in non-linear problems (Kohonen 1982, 2001; Lek *et al.* 1996). These abilities of the Self-Organizing Map can be fully appreciated in view of the goals of this study, i.e., does a group of neurons distinguished by a SOM on the basis of fish abundances confirm or improve groupings of sites differing in qualitative and quantitative species composition as compared to other conventional methods.

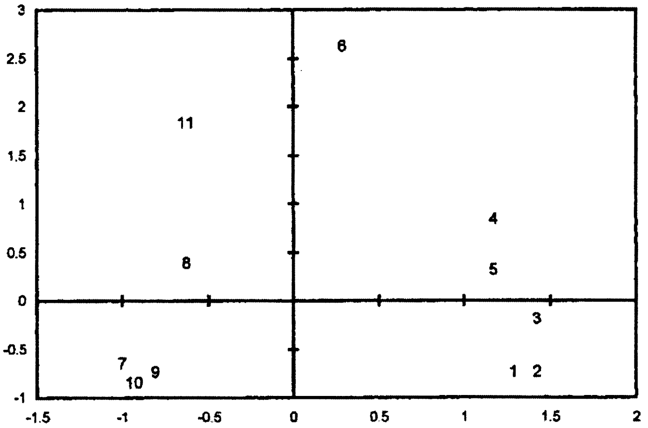

Fig. 5. Diagram of axis one and two for the Canonical Correspondence Analysis relating site scores of the Caracu and Água do Rancho Rivers (after Penczak *et al.* 2002).

The *VUA1* contains three sites (7, 9 and 10) of the Água do Rancho River, which are distinguished by the highest similarity of species composition. This was precisely confirmed by CCA (Fig. 5), though sites linearly ordered by RD (Fig. 6) distort this order somewhat. Between these three sites, site 8 is inserted, which is separated on the SOM and CCA diagrams. However, when we examine the two-dimensional dendrite of the preceding RA diagram (Fig. 6), we can see that site 7 belongs to a group containing sites 10 and 9, but due to its weaker similarity to site 10 than to site 8, it can only be ordered linearly after site 8 (see arrow on the figure). This occurred because when a two-dimensional dendrite is converted to a one-dimensional (linear) one, the principle that the weakest connection must be 'broken' to insert between respective sites those that were not in the linear order (Romaniszyn 1970). The same situation occurred with the distribution of the Caracu River sites 1, 2 and 3. In the SOM, sites 1 and 2 are located in one virtual unit (*VUD1*) adjoined by site 3 in another unit (*VUD2*). Conversely, in the two-dimensional space of CCA, sites 1 and 2 are very close to each other while site 3 is separated from them. In contrast in the linear RD dendrite, sites 1 and 2 are near each other, but site 3 is separated from them by site 4 (Fig. 6). However, in the two-dimensional dendrite, site 3 is immediately connected with site 2, but according to the principle mentioned above, it is inserted after site 4. Such problems exist if the linear order of sites is used instead of multivariate space (Lek *et al.*, 1996; Giraudel, Lek 2001; Park *et al.* 2003). Despite these results, the final synthetic diagram of RD, based on the linear dendrites of sites and species, produces a clear picture that approximates reality

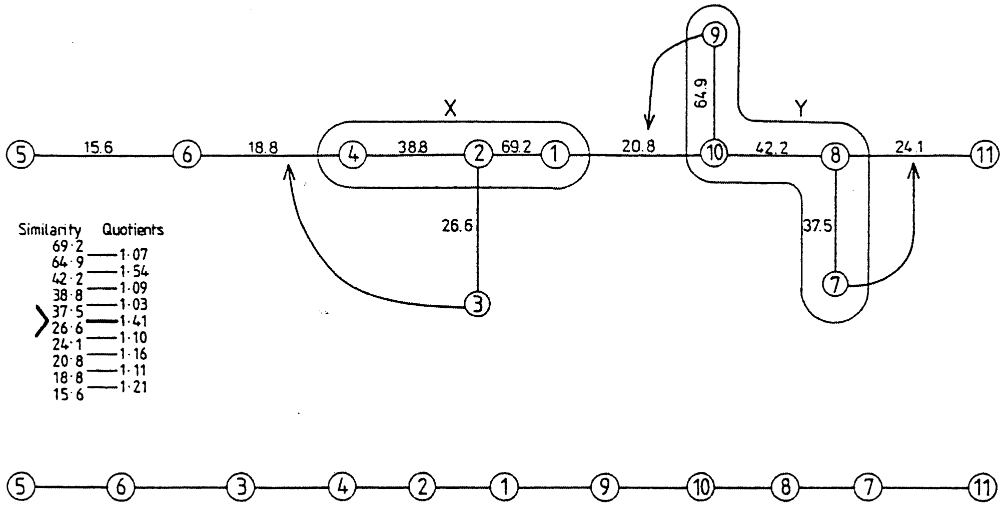

Fig. 6. The two-dimensional dendrite of sites, with two groups of sites (X, Y) distinguished on the basis of a quotient of 1.41. Below it is the linear dendrite (after Penczak *et al.*, 1994).

in the distribution of sites and species (Fig. 7). The synthetic diagram similar to the SOMs with fish species composition, provides useful information on the co-occurrence of species, while the former one has provided precise information on how abundant the individual species are in a given site, and potential density-dependent relationships among species. Of course a respective diagram of CCA, but especially of SOM, have ordered the analysed sites in a more objective way (Lek *et al.*, 1996). However, the application of RD is not connected with long-term study as in the case of gradient analyses or especially the Artificial Neural Network methods. In addition, RD does not promote the improper result interpretations that accompany gradient analysis (Palmer 2000, ter Braak 1987). Despite admiring the RD, we have to state that more complex data sets (more than 50 sites, for example) may limit the application of the method. Apart from the calculation of sites and species similarities and drawing a synthetic diagram, a computer program needed to perform this analysis is not yet available. Transformation of the two-dimensional dendrite by hand to the one-dimensional one is subject to error.

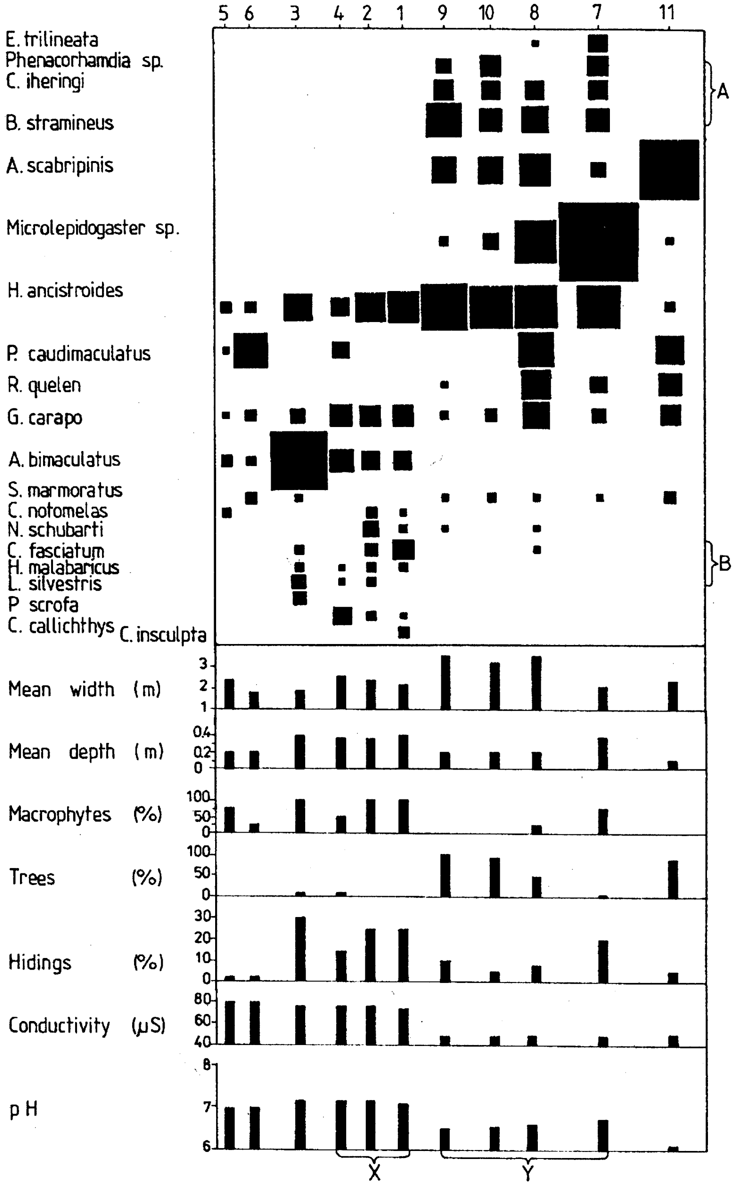

Fig. 7. Synthetic diagram showing quantitative species distribution in relation to sites (1-11). Sites characteristics related to fish population abundance are presented below the diagram. A and B are species clusters; X and Y are site clusters - distinguished on the basis of a quotient calculated between neighbouring similarity values (after Penczak *et al.* 1994).

The Romaniszyn method, which was used by us to examine patterns in fish assemblages and sites, positively passed the verification by the CCA and SOM. We revealed minor weaknesses of the RD, but we still recommend it for scientists unfamiliar with gradient analysis and ANN. However, we have to underline that the SOM algorithm is one of the best techniques in ecology for analysing populations data and especially for community ordination (Chon *et al.* 1996). The method solves difficult high-dimensional and non-linear problems (Kohonen 2001; Lek *et al.* 1996), which are very common in community-level studies.

Acknowledgements

The authors wish to thank the editor of Hydrobiologia, H.J. Dumont, and Kluwer Academic Publishers (Belgium) for permission to reuse published Figure 6 and 7, and the editor of Limnologica, W. Schönborn, and Urban & Fischer Verlag for permission to reuse Figure 5. We thank Ł. Głowacki for improvement of the English, and Andrzej Kruk for drawing original figures. We also thank Stephen Axford and an anonymous referee for final English correction and valuable comment. T. Penczak would like to thank Young-Seuk Park for introducing him to the Kohonen, SOM algorithm.

5. References

Adamczyk, J., Głowacki, Ł., Penczak, T. 2004. Structure of macrofungus communities in different habitats of small postglacial pond margins. *Acta Oecologica* **25**, 53-60.

Agostinho, A.A., Penczak, T. 1995. Populations and production of fish in two small tributaries of the Parana River, Parana, Brazil. *Hydrobiologia* **312**, 153-166.

Brosse, S., Giraudel, J.L., Lek, S. 2001. Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecological Modelling* **146**, 159-166.

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

Chon, T.S., Park, Y.S., Park, J.H. 2000. Determining temporal pattern of community dynamics by using unsupervised learning algorithms. *Ecological Modelling* **132**, 151-166.

Giraudel, J.K., Lek, S. 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling* **146**, 329-339.

Giraudel, J.L., Aurelle, D., Berrebi, P., Lek., S., 2000. Application of the self-organizing mapping and fuzzy clustering microsatellite data: how to detect genetic structure in brown trout (Salmo trutta) populations. In: Lek, S., Guegan, J.F. [Eds] *Artificial Neural Networks: Application to Ecology and Evolution, Environmental Science.* Springer, Berlin, pp. 187-201

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetic* **63**, 201-208.

Kohonen, T. 1995. *Self-Organizing Maps.* Springer Series in Information Sciences. Second Extended Edition. Springer, Berlin.

Kohonen, T. 2001. *Self-Organizing Maps.* Third Extended Edition. Springer, Berlin.

Lek, S., Delacoste, M., Baran, P., Dimopulos, I., Lauga, J., Aulagnier, S. 1996. Application of neural networks to modeling nonlinear relationships in ecology. *Ecological Modelling* **90**, 39-52.

Matthews, W.J. 1998. *Patterns in freshwater fish ecology.* Kluwer Academic Publishers. New York.

Palmer, M. 2000. *Ordination methods for ecologists.* http://www.okstate.edu/artsci/botany/ordinate

Park, Y.S., Kwak, I.S., Chon, T.S., Kim, J.K., Jorgensen, S.E. 2001. Implementation of artificial networks in patterning and prediction of exergy in response to temporal dynamics of benthic macroinvertebrate communities in streams. *Ecological Modelling* **146**, 143-157.

Park, Y.S., Céréghino, R., Compin, A., Lek, S. 2003. Application of artificial neural network for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling* **160**, 265-280.

Penczak, T., Agostinho, A.A., Okada, E.K. 1994. Fish diversity and community structure in two small tributaries of the Parana River, Parana State, Brazil. *Hydrobiologia* **294**, 243-251.

Penczak, T., Agostinho, A.A., Głowacki, L., Gomes, L.C. 1997. The effect of artificial increases in water conductivity on the efficiency of electric fishing in tropical streams (Paraná, Brazil). *Hydrobiologia* **350**, 189-201.

Penczak, T., Agostinho, A.A., Hahn N.S. 2000. An ordination technique for fish diet comparison. *Brazilian Archives of Biology and Technology* **43**, 101-110.

Penczak, T., Agostinho, A.A., Latini, J.D. 2003. Rotenone calibration of fish density and biomass in a tropical stream sampled by two removal methods. *Hydrobiologia* **510**, 23-38.

Penczak, T., Godinho, F., Agostinho, A.A. 2002. Verification of the dualism ordering method by the canonical correspondence analysis: fish community samples. *Limnologica* **32**, 14-20

Romaniszyn, W., 1970. Próba interpretacji tendencji skupiskowych zwierząt w oparciu o definicję podobieństwa i odległości [An attempt at interpreting agglomerative tendencies of animals based on definition of similarity and distance]. *Wiadomości Ekologiczne* **16**, 306-327 [Engl. summ.].

ter Braak, C.J.F. 1987. Ordination. In: Jongman, R.H.G., ter Braak, C.J.F., van Tongeren, O.F.R. [Eds] *Data analysis in community and landscape ecology,* Wageiningen, Pudoc, pp. 91-173.

ter Braak, C.J.F. 1990. *CANOCO version 3.1.* Update notes. Wageningen, Netherlands: Agricultural Mathematics Group.

ter Braak, C.J.F., Verdonschot, P.F.M. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Science* **57**, 255-289.