Available online at www.sciencedirect.com

SCIENCE DIRECT®

Science of the Total Environment 327 (2004) 105–122

Science of the Total Environment
An International Journal for Scientific Research into the Environment and its Relationship with Humankind

www.elsevier.com/locate/scitotenv

Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks

Young-Seuk Park[a,]*, Tae-Soo Chon[b], Inn-Sil Kwak[b], Sovan Lek[a]

[a]*LADYBIO, CNRS-University Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, France*
[b]*Division of Biological Sciences, Pusan National University, Busan 609-735, South Korea*

Accepted 13 January 2004

Abstract

Benthic macroinvertebrate communities in stream ecosystems were assessed hierarchically through two-level classification methods of unsupervised learning. Two artificial neural networks were implemented in combination. Firstly, the self-organizing map (SOM) was used to reduce the dimension of community data, and secondly, the adaptive resonance theory (ART) was subsequently applied to the SOM to further classify the groups in different scales. Hierarchical grouping in community data efficiently reflected the impact of the environmental factors such as topographic conditions, levels of pollution, and sampling location and time across different scales. New community data not included in the training process were used to test the trained network model. The input data were appropriately grouped at different hierarchical levels by the trained networks, and correspondingly revealed the impact of environmental disturbances and temporal dynamics of communities. The hierarchical clusters based on a two-level classification method could be useful for assessing ecosystem quality and community variations caused by environmental disturbances.

Keywords: Self-organizing map; Adaptive resonance theory; Two-level community classification; Benthic macroinvertebrates; Multivariate analysis

1. Introduction

Community patterning is useful for revealing ecological states of ecosystems in response to environmental disturbances. In aquatic ecosystems, community compositions vary rapidly against stressful sources of natural and anthropogenic origins such as flooding, pollution, etc. (Hawkes, 1979; Hellawell, 1986; Spellerberg, 1991). Classification and ordination of communities have been recently focused on water quality assessment. Among different taxa, benthic macroinvertebrate communities are effective in indicating water quality and could effectively reveal ecological states of aquatic ecosystems (Hynes, 1960; Hawkes, 1979; Hellawell, 1986). They constitute a heterogeneous assemblage of animal phyla, and consequently it is probable that some members will always respond to stresses placed upon them.

*Corresponding author. *Present address*: Research Unit of Water Quality, Cemagref Bordeaux, 50 avenue de Verdune, 33612 Cestas, France. Tel.: +33-5-57-89-09-90; fax: +33-5-57-89-08-01.

E-mail address: park@cict.fr (Y.-S. Park).

0048-9697/04/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2004.01.014

Community data, however, are non-linear and complex: they consist of many species, being highly variable in densities affected by various biotic (e.g. physiological development, life cycle, etc.) and abiotic factors (e.g. precipitation, pollution, etc.) (Jongman et al., 1995; Legendre and Legendre, 1998). Recently, biological indicator systems have been developed based on analyses of community data. In aquatic ecosystem management, for instance, the River Invertebrate Prediction And Classification System (RIVPACS) has been proposed for assessing the biological quality of freshwater. The RIVPACS and its derivatives have been the primary ecological assessment analysis techniques for Great Britain (Wright et al., 1993) and Australia (Norris, 1995). Based on a stepwise progression of multivariate and univariate analyses, the models predict the aquatic macroinvertebrate fauna that would be expected to occur at a site in the absence of environmental stress (Barbour et al., 1999; Coysh et al., 2000). These models are, in general, based on conventional multivariate statistical methods (Ludwig and Reynolds, 1988; Jongman et al., 1995; Legendre and Legendre, 1998). Among them, ordination techniques were used to obtain a way to display statistical sample units, which are considered to be drawn from a population whose variations are continuous (Goodall, 1954; Giraudel and Lek, 2001). Diverse linear ordination methods have been implemented for compressing data size, for example, polar ordination, principal components analysis (PCA), correspondence analysis (CA), etc. (Pearson, 1901; Hill and Gauch, 1980; Beals, 1984; Jongman et al., 1995).

The limitations of the conventional methods, however, are well known: strong distortions with non-linear species abundance relations (Kenkel and Orloci, 1986), horseshoe effect due to unimodal species response curves in PCA, disjointed data matrix in CA, arch effect, outliers, missing data, etc. (Giraudel and Lek, 2001). Recently, as an alternative tool to deal with this problem of complexity in ecological data, Kohonen's self-organizing map (SOM) (Kohonen, 1982, 2001) has been used for patterning samples in diverse ecosystems (i.e. aquatic, forest, agriculture, etc.) (Lek and Guégan, 2000; Recknagel, 2002): for community classification (Chon et al., 1996, 2000; Park et al., 2001, 2003a), water quality assessments (Walley et al., 2000; Aguilera et al., 2001), prediction of population and communities (Céréghino et al., 2001; Obach et al., 2001), and conservation strategies of endemic species (Park et al., 2003b). Classification of communities by the SOM, however, encounters a problem of objectivity in finding similarities among the classified samples (Chon et al., 1996). The SOM produces a map of classification in a low dimensional lattice (commonly two-dimensional) consisting of computational output units. When the groups are located far apart on the map, it is difficult to judge to what extent they are similar. Furthermore, due to randomness in iterative calculations and variability in determining parameters in learning processes, the grouping presents a slightly different conformation after each training task. In this study, we propose a combinational method of supervised learning to alleviate the problem of objectivity in grouping and to demonstrate the feasibility of hierarchical classification for assessment of aquatic ecosystems.

2. Materials and methods

2.1. Ecological data

The benthic macroinvertebrate community data were provided by the Laboratory of Ecology and Behavior Systems, Pusan National University, Korea. The data were seasonal samples collected at the sites across different levels of pollution in the Suyong (SY), Cheolma (CM), Hoedong (HD), and Soktae (ST) streams in the Suyong River in Korea (Fig. 1a) in October, 1989, and in January, May, and August, 1990. The Suyong River is a fourth order river, 28.5 km in length with a catchment area of 199.5 km^2, passing through the Pusan city area. Two tributaries Cheolma and Suyong flow through agricultural areas to the Hoedong reservoir. The Hoedong, which is located in the lower area of the reservoir, is characterized by abundant filamentous algae and low current velocity, but with a great variation in discharge rates due to the water drained from the reservoir. The Soktae stream runs through the populated

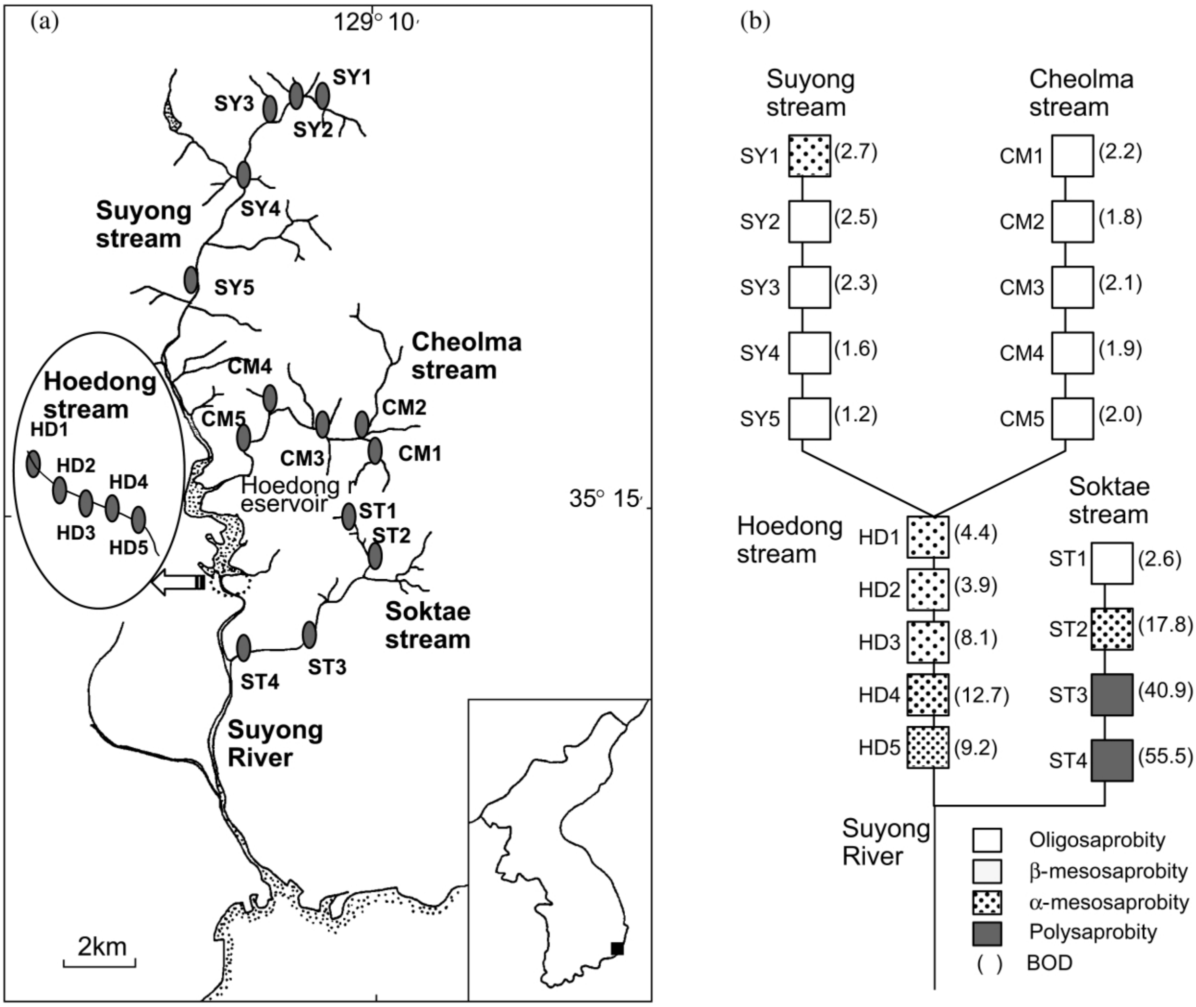

Fig. 1. Hydrological network of rivers studied and location of sampling sites (a) and their pollution states (b).

residential area, being heavily polluted by organic matter in domestic sewage (Kwon and Chon, 1991, 1993).

Benthic macroinvertebrates were collected using a Surber sampler (30×30 cm). The dataset consisted of 76 samples from 19 sites in four seasons. Eighty-four species were recorded in datasets in total. Dominant taxa were Chironomidae, Tubificidae, Erpobdellidae, Hydropsychidae and Baetidae. In Cheolma, 38 species including *Ordobrevia* sp. and *Parachauliodes continentalis* were identified, in Suyong 34 species including *Enchytraeus* sp. and *Synorthoccladius* sp., in Hoedong 40 species including *Cardina dentriculata* and *Helobdella* sp., and in Soktae 17 species including *Chironomus* sp. and *Limnodrilus hoffmeisteri*.

The Biological Monitoring Working Party (BMWP) score (Walley and Hawkes, 1996, 1997; Hawkes, 1997) was obtained from the sampled data as a biological water quality index, showing high variations at different sampling sites with mean values of 57.7 (range 36.1–88.6) in Cheolma, 39.3 (range 12.7–73.1) in Suyong, 54.4 (range 20.6–99.1) in Hoedong, and 14.6 (range

7.2–41.4) in Soktae. A wide range in organic pollution was observed in the study area (Fig. 1b), showing from oligosaprobity to polysaprobity. Community structure and ecological assessment on the Suyong River have been reported in Kwon and Chon (1993) and Chon et al. (2000).

The abundances of 84 species were provided as input to the network (number of computation nodes in the input layer = 84). The data were transformed by natural logarithm in order to reduce variations in abundance. To avoid problems of logarithm zeros, the number one was added to the density of each species. Subsequently the transformed data were proportionally scaled between 0 and 1 in the range of the minimum and maximum density for each species. Normalizing the density data is necessary because if one variable has values in the range of large differences, the biggest will tend to dominate the organization of the SOM map because of its greater impact on the distances (Euclidean distance in this case) measured. The standard way to achieve this is to linearly scale all variables (Vesanto et al., 1999; McCune et al., 2002). After the learning processes, a new dataset, which was seasonally sampled at SY2 (see map in Fig. 1) from 1993 to 1995, was provided to the network for validation.

2.2. Modeling process

The two-step classification processes with unsupervised learning were applied to community data. First, the densities of different taxa of benthic macroinvertebrate communities were fed into the SOM. After training, weight vectors (i.e. connection intensities) of the SOM, containing conformational characteristics of grouping in communities, were subsequently provided to another self-organizing network, the adaptive resonance theory (ART; Carpenter and Grossberg, 1987), to find clusters in the units of the SOM. Different levels of dissimilarity threshold in the ART were provided for clustering in different scales. Both the SOM and the ART are used for clustering datasets. However, the main difference concerns topology preservation of output units of networks: neighboring topology is preserved in SOM, while it is not preserved in ART. According to the topology preservation of the SOM, the SOM is more preferred in the ordination of samples than the ART.

2.2.1. Self-organizing map (SOM)

In the SOM learning process, initially the weight vectors are randomly assigned small values. When the input vector x is sent through the network, the distance between the weight vector w and the input vector x is calculated by Euclidean distance $||x-w||$. The output layer consisted of N output nodes (i.e. computational units) on a two-dimensional hexagonal lattice (Kohonen, 2001). Among all N output units, the best matching unit (BMU), which has the minimum distance between weight and input vectors, becomes the winner. The weight,w_{ij}, of the network is updated as follows:

$$w_{ij}(t+1)=w_{ij}(t)+\eta(t)N(t,r)\left[x_j-w_{ij}(t)\right] \quad (1)$$

where $\eta(t)$ denotes the fractional increment of the correction, and $N(t, r)$ is a predefined neighborhood function determining the radius from the BMU in the map. The neighborhood radius was usually set to a larger value early in the learning process, and was gradually reduced on approaching convergence. The detailed algorithm of the SOM can be found in Kohonen (2001) for theoretical considerations, and Chon et al. (1996), Park et al. (2001, 2003a) and Lek and Guégan (2000) for ecological applications.

The map size (number of output units) of the SOM is critical for accommodating hierarchical levels in community classification. We trained the SOM with different map sizes, and chose the optimum map size based on the minimum values of quantization and topographic errors. The quantization error is the average distance between each input vector and its BMU and is used to measure map resolution (Kohonen, 2001). The topographic error represents the accuracy of the map in preserving topology; the error value is calculated from the proportion of all data vectors for which first and second BMUs are not adjacent for measuring topology preservation (Kiviluoto, 1996).

During the learning process, nodes that are topographically close in the array will activate each other to establish coherence from the same input vector. This results in a smoothing effect on the weight vectors of nodes (Kohonen, 2001). To analyze the contribution of variables to cluster structures of the SOM, each input variable (species) calculated during the training process was visualized in each node on the trained SOM in gray scale. Based on the component planes, the correlation coefficients were calculated between component pairs in both observed and calculated data. We used the functions implemented in the SOM toolbox (Alhoniemi et al., 2000) for Matlab (The Mathworks, 2001) developed by the Laboratory of Information and Computer Science in the Helsinki University of Technology. We adopted the initialization and training methods suggested by Alhoniemi et al. (2000) that allow the algorithm to be optimized. The software library is available from the website http://www.cis.hut.fi/projects/somtoolbox.

2.2.2. *Adaptive resonance theory* (*ART*)

The ART is capable of achieving stable self-organization of datasets for an arbitrary number of input vectors (Carpenter and Grossberg, 1987). The fundamental characteristic of the ART lies in its ability to dynamically self-adjust its output size depending on the complexity of the network (Baraldi and Alpaydin, 1998). The algorithm selects the first input as the exemplar for the first cluster, and new input is subsequently clustered with the first if the distance to the first is less than a given threshold. It plays the role of the exemplar for a new cluster. A special parameter residing in the ART is the threshold value for determining vigilance (Lin and Lee, 1996). As the threshold value increases, the group size increases accordingly. This provides a basis for organizing the input data in different hierarchical levels.

A modified algorithm in ART (Pao, 1989) was used in this study. Bottom up weights $b_{lk}(1)$ between input node l and output node k are initialized with some small numbers. After the input value y_l, which is the weight of each output node of the SOM, has been fed into the network, the distance $d_k(t)$ is measured for the degree of dissimilarity between input and weight values for each output node k, and used as a criterion for grouping input data through training.

The weight vectors produced from the SOM were fed into the ART. As each new input vector is sent into the ART, the distance is calculated and the output node k which is closest to the new input is selected as k^*. If $d_{k^*}(t)$ is smaller than ρ, which is a threshold parameter for determining vigilance, the input vector is assigned to output node k^*. The weight of node k^*, $b_{lk^*}(t)$, is then updated as follows:

$$b_{lk^*}(t+1)=\frac{c}{c+1}b_{lk^*}(t)+\frac{1}{c+1}y_l \tag{2}$$

where c is the number of sample units classified to node k^*. If $d_{k^*}(t)$ is larger than ρ the input is assigned to a new output node. This means that the input vector entered is ‘patterned’ (or classified) as a new pattern (or cluster), not belonging to one of the patterns existing previously. Then, its weight $b_{lk^*}(t)$ is newly assigned with the input vector. For hierarchical clustering in this study, initially the learning process was begun with whole map sizes. When clusters were found, the input vectors were replaced with their corresponding weights from the ART. The ART program was developed to run under the Matlab (The Mathworks, 2001) environment.

2.2.3. *Unified-matrix algorithm* (*U-matrix*)

To compare the clustering abilities of the ART for the SOM units, U-matrix algorithm (Ultsch, 1993), a commonly-used quantification method for finding associations between SOM units, was also applied in this study. The U-matrix calculates the distance of a weight vector (w) to its neighbors in the SOM, and displays the cluster structure of the map units. Supposing the map has a size of m columns and n rows, the following value ($M_{x,y}$; U-matrix) is calculated for all positions:

$$M_{x,y}=\frac{1}{H}\sum_{a=x-1}^{x+1}\sum_{b=y-1}^{y+1}\left\|w_{x,y}-w_{a,b}\right\| \tag{3}$$

where H is the number of neighbor units ranging from 2 to 6, depending upon the location of the map unit. The values were rescaled between 0 and 1 for the purpose of visual comparison. The matrix

Table 1
Changes of quantization error and topographic errors at different SOM map sizes

Map size	10	20	30	40	54	63	70	80
Quantization error	1.561	1.476	1.382	1.355	1.290	1.256	1.242	1.234
Topographic error	0.000	0.013	0.000	0.000	0.000	0.000	0.001	0.000

was presented as a gray-scaled picture based on the calculated values: bright areas with low values depict short distances while dark areas with high values represent long distances to the surrounding neighbors. High values of the U-matrix indicate groups' boundaries, while low values reveal groups themselves.

2.2.4. Principal component analysis (PCA)

The visualization method of the SOM could be related to PCA. To compare the ordination performance of SOM, PCA was carried out with the same datasets used for SOM training using a statistical package STATISTICA (Statsoft, 2000).

3. Results

3.1. Benthic community patterns

The number of output units of the SOM is important to classify communities as stated previously. If the map size is too small, it might not explain some important differences that should be detected. Conversely, if the map is too big, the differences are too small. After preliminary tests with two different indices (quantization and topographic errors) with different map sizes, we chose 63 (9×7) output units of the SOM on the two-dimensional hexagonal lattice (Table 1) in lower errors in both indices. The errors for quantization and topography were 1.256 and <0.001, respectively. Quantization error, representing average distance between each input vector and its BMU, decreased gradually according to the reduction of the map size. One SOM unit has a higher possibility of being occupied by only one input vector in a larger map size than in a smaller one. Therefore, the larger the map size, the lower the errors observed between input vector and output vector. The values of topographic error appeared to be very low in all cases in this study, indicating that the first and second BMUs of all input vectors were adjacent hexagons and smooth training was conducted in the SOM.

The trained SOM classified samples according to variations and gradients observed in benthic macroinvertebrate communities (Fig. 2). The acronyms in each unit of the SOM map stand for the samples. The first three letters in the acronyms indicate the study sites (Fig. 1), while the last three characters represent the sampling seasons: SPR; spring, SUM, summer, AUT; autumn, and WIN; winter (e.g. ST1SPR; samples at ST1 in spring). The grouping was firstly arranged according to the geographical distribution of the sample sites, e.g. Soktae, Hoedong, Cheolma and Suyong. The samples collected from Cheolma are mostly located in the lower area of the SOM, while those belonging to Hoedong are more concentrated in the upper right area. Furthermore, temporal variations in different seasons were also observed locally. For instance, the samples collected in summer at SY1-5 were either grouped together in the same unit or were located near each other (e.g. nodes (5 (row), 3 (column)) and (6, 2)).

The grouping on the map also revealed the impact of pollution, being comparable with pollution states of the sampling sites (Fig. 2). The upper area of the trained map represented polluted sampling sites, whereas the lower area showed relatively clean sites. For example, sites ST2-4 were concentrated in the upper left corner of the SOM (e.g. nodes (1,1), (2,1) and (3,1)). The sites were heavily polluted from domestic sewage, showing polysaprobity. Moreover, the samples from the less polluted site, ST1, with oligosaprobity, were not strongly grouped with sites ST2-4 and were more or less widely scattered in the upper left area of the map. The sites HD4-5, heavily affected by domestic and industrial wastes, formed another strong group at the top of the map near the area of ST2-4. The less polluted sites

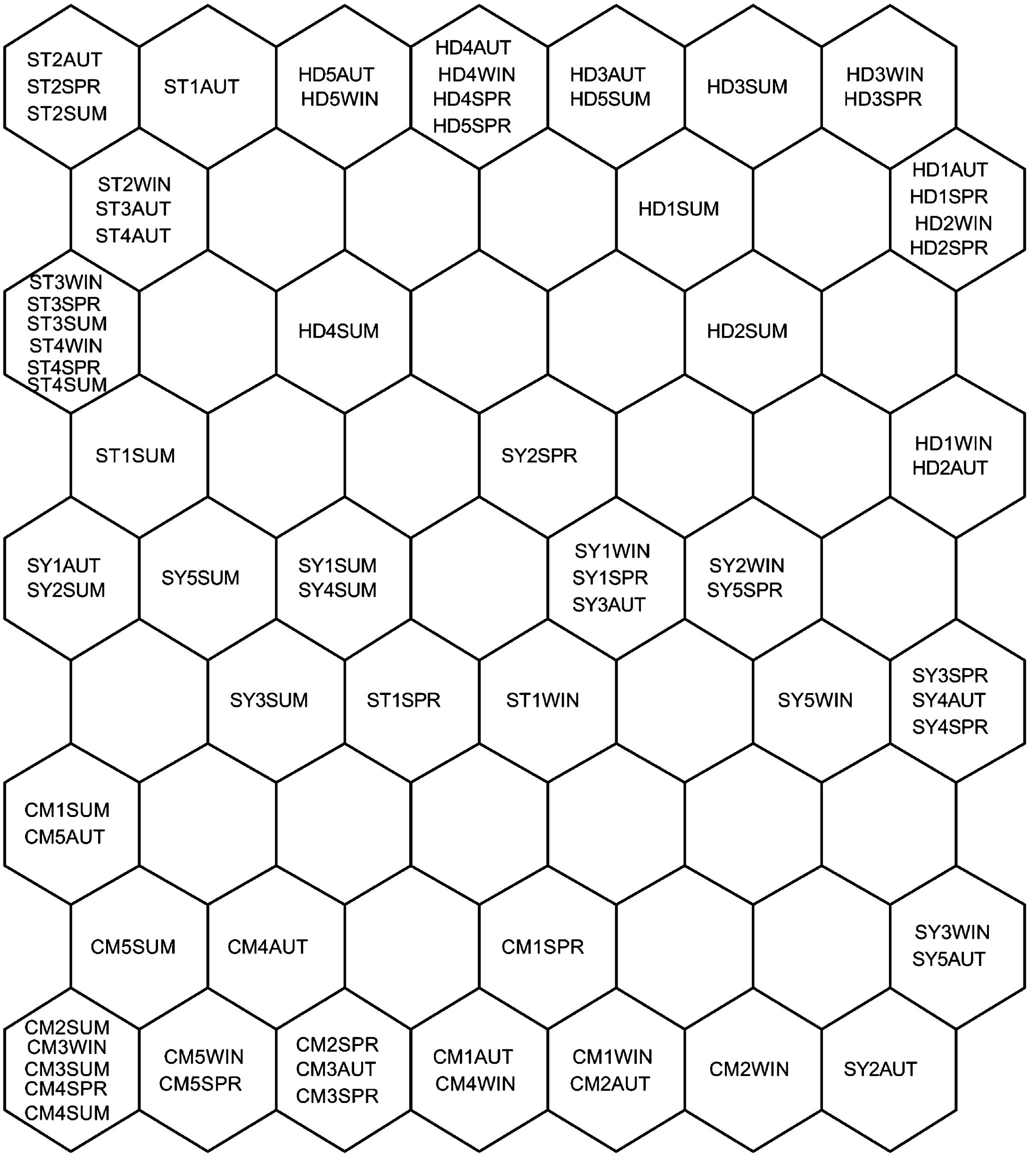

Fig. 2. Classification of sampling units by the trained SOM. Acronyms in units stand for samples: the first three letters represent sampling sites (see Fig. 1), and the last three indicate the sampling season: SPR; spring, SUM, summer, AUT; autumn, and WIN; winter.

HD1-2 were placed more loosely in the upper right area of the map. Site HD3 was relatively clean but occasionally disturbed by domestic waste. The samples of HD3 fell on the boundary between the areas of HD4-5 and HD1-2. The clean sites from Cheolma with oligosaprobity (Fig. 1b) were mostly located in the lower area of the SOM. The sampling sites of Suyong were relatively clean ranging from oligosaprobity to β-mesosaprobity, and communities from these sample sites were diverse. Correspondingly, the samples occupied a wide range in the middle area of the map.

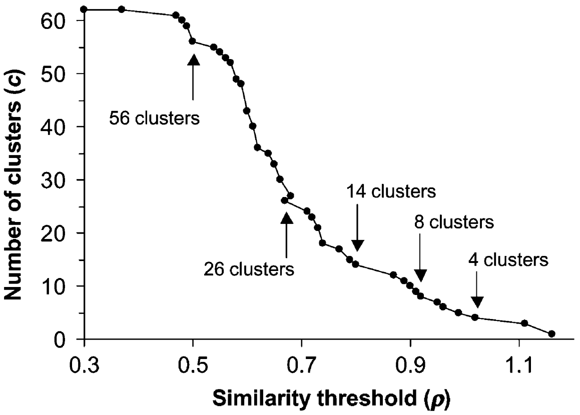

Fig. 3. Relationships between number of clusters in the SOM units and dissimilarity thresholds in the ART. The network was trained at different dissimilarity threshold levels to find clusters of the SOM units.

3.2. Hierarchical classification using ART

Although the samples were grouped on the SOM map, it is still difficult to recognize the differences among the units of the SOM map. In this study we produced additional clusters in different hierarchical levels by implementing the ART. For this purpose, the weights of the trained SOM were provided to the ART as input, and the units of the SOM map were further grouped in different scales by adjusting dissimilarity threshold levels in the ART. The number of input nodes was 84; each node corresponds to one species used in the SOM.

As the dissimilarity threshold was increased, the number of groups decreased correspondingly (Fig. 3) to reveal higher levels in the hierarchy. The optimal number of clusters can be determined based on the relationships between the number of clusters (c) and the threshold values (ρ): the number of clusters was stabilized (i.e. $\mathrm{d}c/\mathrm{d}\rho = 0$ and $c \neq 1$) at various points as dissimilarity threshold values were gradually increased. In this study three stabilized levels of dissimilarity were mainly observed in higher ranges (Fig. 3): 14 subclusters in the range 0.79–0.87 of similarity threshold, eight medium-sized clusters in the range 0.91–0.95, and four large clusters in the highest level in the range 0.99–1.11. Small groupings with 26 and 56 clusters were also observed at lower similarity thresholds (Fig. 3), however, grouping at these low levels was too detailed for differentiation, and were not considered in the present study.

Fig. 4 shows the hierarchical clustering projected on the SOM map. With the low number of clusters at the higher dissimilarity level (Fig. 4a), the samples fell into four clusters (I–IV): the groups of Soktae–Hoedong–Suyong (cluster I), Cheolma (cluster II), Hoedong (cluster III), and Suyong (cluster IV) (Fig. 4a). The groupings reflected the differences in sampling locations and pollution levels. The group Soktae–Hoedong–Suyong (cluster I) accommodated a wide range of sample sites with pollution sites in the upper left area and the intermediately polluted sites of Suyong in the middle area of the SOM. The communities in Suyong were diverse in particular, and were divided into two different clusters with different levels of pollution, Suyong (cluster IV) and Soktae–Hoedong–Suyong (cluster I). The sample sites in Hoedong also showed different levels of pollution. Sites HD4-5 were located in cluster I (Soktae–Hoedong–Suyong group), showing α-mesosaprobity, while the less polluted sites HD1-3 below the Hoedong reservoir joined cluster III in β-mesosaprobity. The relatively clean sites from Cheolma (cluster II) occupied the lower area of the SOM map.

The eight medium-sized clusters (Ia–IV) in the lower hierarchical level were also classified based on the pollution states and the streams (Fig. 4b). Large cluster I was divided into four subgroups Ia, Ib, Ic and Id according to pollution levels. Cluster Ia showed groups of the highly polluted sites, including ST3-4 with polysaprobity. Cluster Ic mostly accommodated HD4-5 with α-mesosaprobity, while the remaining groups showed relatively lower levels of saprobity. Unlike cluster I, the communities collected from the same stream in cluster II were separated based on the stream gradient: upstream CM1-2 (cluster IIb) and downstream CM4-5 (cluster IIa). Clusters III and IV, which occupied smaller areas in the SOM, were not divided in the medium sized clustering (Fig. 4b).

At the lower hierarchical level with fourteen clusters (Fig. 4c), the sample sites were further divided depending upon the impact of various

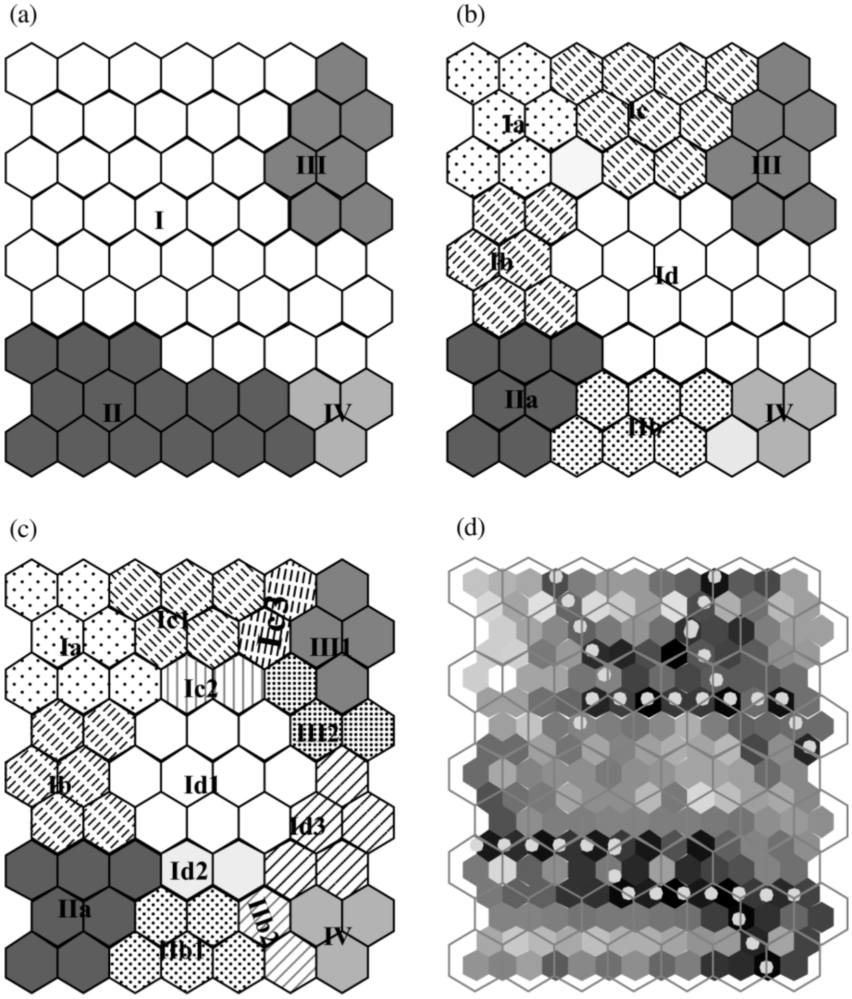

Fig. 4. Hierarchical clustering of the trained SOM map by the ART and clustering by the U-matrix. The different clusters are displayed with characters. From four large clusters (a), the clusters were divided into subclusters based on the corresponding similarity levels (b, c). The symbols represent the classification directions of each cluster. I–IV stand for the four large clusters (a), the eight medium-sized clusters (b), and the 14 small clusters (c). The clusters determined by the U-matrix are indicated with white dotted lines based on the gray scale of U-matrix distance (d).

environmental factors including location of sample sites, season and other disturbances. For example, communities in cluster Id in the middle area of the map were further sub-subgrouped according to the locations of SY1-2 and SY4, whereas some communities in subcluster Ic were separated based on season (e.g. Ic3). The sample sites in cluster III were divided into two sub-subclusters III1 and III2. The samples collected in spring and in HD3 were more selectively grouped on sub–subcluster III1. The samples in cluster IV were not sub–subgrouped at the lowest hierarchical level.

3.3. Comparison with U-matrix method

U-matrix was applied to the results of the SOM to be compared with the performance of grouping by the ART (Fig. 4d). High values of the U-

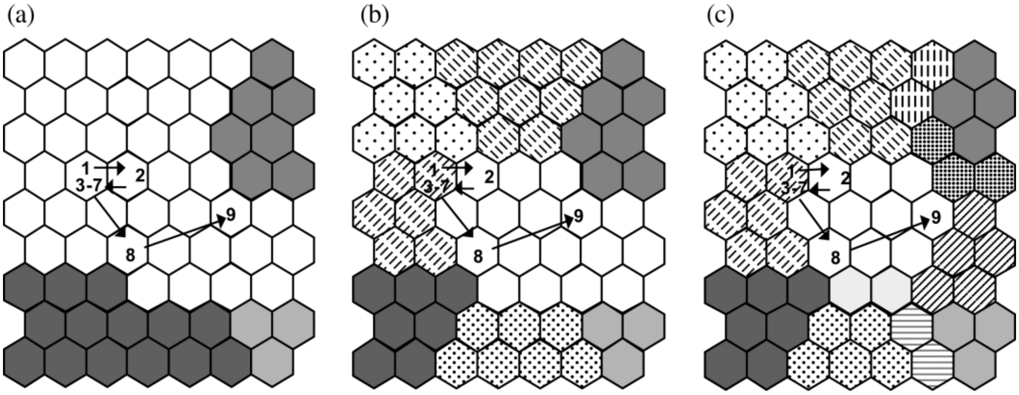

Fig. 5. Evaluation of new communities not used in the learning process at different similarity levels. The data were seasonally collected at SY2 of Suyong stream for 2 years from November 1993 to November 1995. The numbers stand for samples from autumn 1993 to autumn 1995 sequentially. (a) four clusters, (b) eight clusters, and (c) 14 clusters.

matrix indicate cluster boundaries, while low values reveal clusters themselves, which can be visualized in gray scales. The lighter the gray scales between the map units is, the smaller is the relative distance between them. On the U-matrix the nodes of the SOM tended to be grouped similarly to the clustering by the ART at the highest hierarchical level (Fig. 4a). The boundary at the lower area between cluster I and clusters II–IV occurred on the U-matrix, and the boundary between cluster I and cluster III at the upper right area of the map was also observed on the U-matrix. In the upper left area, however, some differences were observed. For instance, the borderline was additionally formed in the U-matrix (Fig. 4d). In the corresponding areas in cluster I on the ART (Fig. 4a), the border line was not observed at the highest hierarchical level. At the subcluster level (Fig. 4b), however, the boundary of cluster Ic was matched to the borderline appearing in the U-matrix. Clustering on the U-matrix, in general, did not appear as sharply as on the ART in general. For example, the sample sites of Soktae were not clearly differentiated from the sample sites of Suyong on the U-matrix (Fig. 4d). Community compositions in actual data were different between Soktae and Suyong. The ART showed more distinctive clusters between Sokae and Suyong at the sub- and sub–sub-grouping levels (Fig. 4b, c). This indicated that the boundaries defined by the ART could more efficiently contribute to clustering of communities than the boundaries on the U-matrix.

3.4. Validating new samples

Once the networks have been trained with input data, new data sets not used for learning can be tested on the SOM; they may be classified either as one of the already determined patterns or as a new pattern at the corresponding hierarchical levels. Through validation, diagnostics of community changes were possible on the hierarchical map. Fig. 5 shows an example of detecting changes in community development in different seasons. The data were seasonally collected at SY2 for 2 years from November 1993 to November 1995. The numbers (1–9) in the nodes of the SOM map stand for sampling seasons from autumn 1993 to autumn 1995 sequentially (e.g. 1 for autumn in 1993, 2 for winter in 1993, etc.). The tracks of the samples recognized represent development of communities in different seasons. At the highest hierarchical level, all the samples belonged to cluster I (Fig. 5a). Consequently community changes were not distinguished at the highest level. However, cross-bordering was observed at subcluster level (Fig. 5b). Changes in the clusters were shown between ‘1 and 3–7’ and other samples. Communities in cluster Ib jumped to cluster

Id1 in winter 1993 and returned to cluster Id1 in summer 1995 (Figs. 2 and 5b). The tracking on the map efficiently elucidates changes across hierarchical levels in community development as time progresses. Cross-bordering in the clusters also reflected differences in water quality. BMWP scores were 30.6 in cluster Ib and 38.55 in cluster Id. The changes in communities were in accord with field observations. The motorway construction and the restoration project of the river had been started since the early 1990s, and were completed sometime early in 1995. Water quality has correspondingly improved since the mid nineties. The jump in the community development was not further detected at the low level of 14 subclusters (Fig. 5c).

3.5. Contribution of species in clustering

An evaluation of the contribution made by input variables (species) to community grouping was possible on the hierarchical map. The approximate probability density function of input data was calculated through an unsupervised learning algorithm in the SOM. Visualization is an efficient way to comprehensively understand the contribution of each species in clustering. Fig. 6 displays component planes of typical distribution patterns of species in the SOM units on a gray scale. Dark represents high abundance of each species, and light shows low abundance. Species showing similar distribution patterns were grouped together according to their distribution gradients. Distribution patterns of species were agreed well with the level of eight clusters (Figs. 4 and 6). Species *Endochironomus* sp. and *Pscycopda* sp. were dominant in cluster Ia on the SOM map, 8 species including *Paraleptophlebia chocorata* were in cluster Ib, 7 species including *L. hoffmeisteri* and *Physa* sp. were in cluster Ic, 12 species including *Cricotopus* sp. and *Hydropsyche* sp. in cluster Id, 10 species including *Psephenoides* sp. in cluster IIa, 15 species including *Ecdyonurus* sp. and *Serratella setigera* in cluster IIb, 16 species including *Asellus higendorfi* and *C. entriculata* in cluster III, and finally 14 species such as *Arctopsyche* sp. and *Batrachobdella* sp. in cluster IV (Fig. 4). Most species showed a clear abundance gradient in the SOM map. Some species, however, did not display clear gradient because they showed high abundance in more than two clusters. For example, *Copera annulata* displayed the highest contribution in cluster Ic, and also in clusters IIb and III. *Orthetrum albistylum* showed the highest values in cluster III and was also abundant in cluster Ic and IIb. Species showing high contribution in more than two clusters were assigned to the cluster displaying the highest value of the SOM weights in Fig. 6.

3.6. Comparison of ordinations between SOM and PCA

In the PCA plot, as in the SOM map, benthic communities were grouped according to degree of association in species composition (Fig. 7). Ordination by the first factor mainly reflected pollution states of sampling sites. The samples from heavily polluted areas were collectively located in the lower left area of the plot (ST3, 4, and HD4, 5). These samples were separately matched to clusters Ia, Ib, and Ic in the SOM map (Figs. 2 and 4), and grouping appeared more clearly in the SOM. In addition to pollution levels, geographical variation of the sampling sites was also observed in the PCA plot along the *y*-axis. The samples from Cheolma (CM1-5) were densely located in lower right area. In contrast, the samples from HD1, 2 and SY3, 5 were located close together in the upper right area. The groups in the PCA correspondingly appeared in the groups shown on the SOM (Figs. 2 and 4). The group of Cheolma (CM1-5) was observed on cluster II on SOM, while the groups of HD1, 2 and SY3, 5 were matched to clusters Id and III in the SOM. In general, however, the samples were more distinctively clustered in the SOM than in the PCA. For instance, the samples classified in clusters in Id and IV in the SOM were not grouped clearly, but scattered over a wide area in the PCA (Fig. 7). In the PCA, seasonal variations in the samples were not clearly observed, although some local samples (i.e. SY1 and SY4 in spring, SY2, 3, 5 in winter) displayed seasonality (Fig. 7). Clusters defined by the SOM, however, showed temporal variations

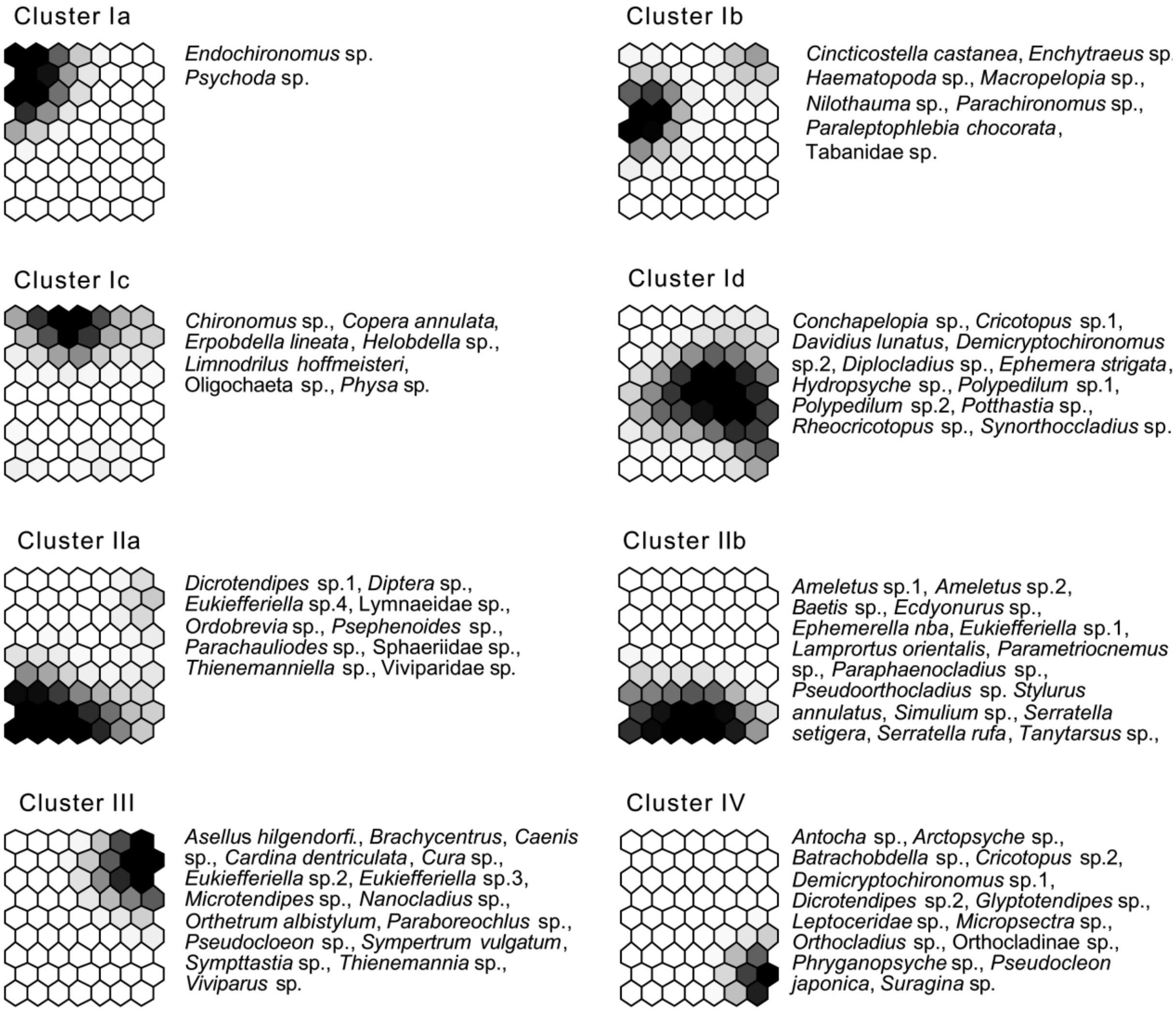

Fig. 6. Typical distribution patterns of species in each cluster in the trained SOM. Dark indicates high abundance of species, and light is low values. Some species made a high contribution to more than one cluster. They were assigned to the cluster displaying the highest contribution.

more clearly, along with pollution states and with spatial location of the streams.

4. Discussion and conclusion

The adaptive learning algorithms efficiently extract information from complex, multivariate community data, and provide a comprehensive understanding of ecological states of ecosystems in a reduced dimension (Chon et al., 1996; Lek and Guégan, 2000; Recknagel, 2002; Park et al., 2003a,b). In this study we designed a type of hybrid network that combines the SOM and the ART. The main advantage of this approach is that it enables samples to be clustered at different similarity levels on the SOM and was able to visualize the samples and species in a reduced dimension in a hierarchy. Through the SOM, communities were classified based on similarities of community composition (Fig. 2). When the ART

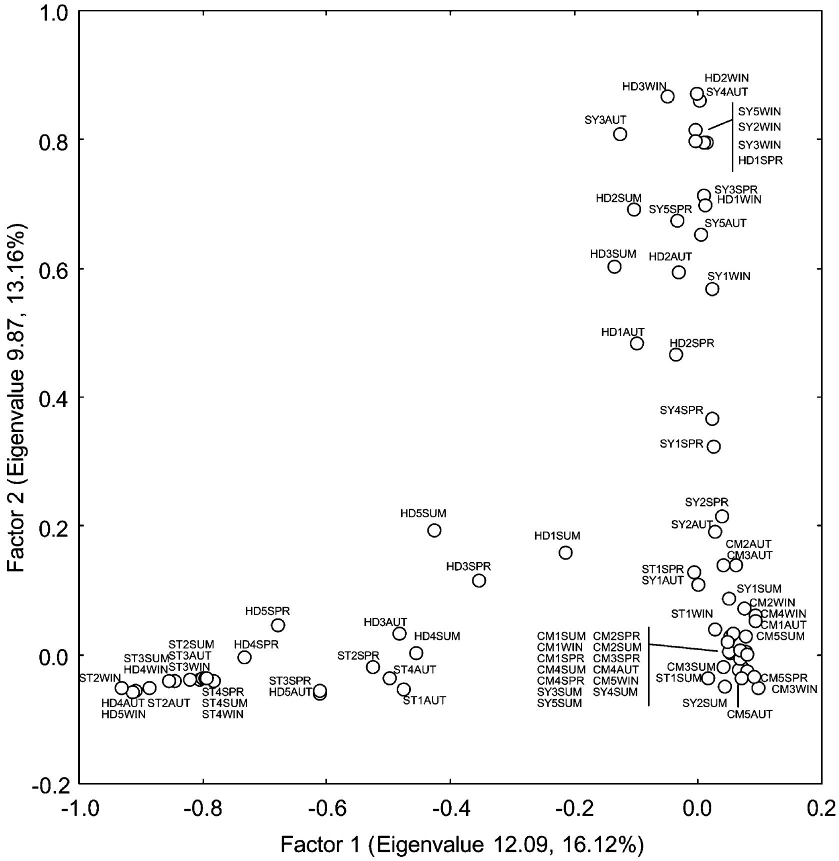

Fig. 7. Ordination of PCA with benthic macroinvertebrate communities. The ordination is compared to the visualization of the SOM in Fig. 2. The first principal factor showed 16.12% of total variance (eigenvalue 12.09), whereas the second one was 13.16% (eigenvalue 9.87).

was further applied to the groupings in the SOM, hierarchical clusters were formed at different dissimilarity levels (Fig. 4a–c).

The SOM approximates the probability density function of input data through an unsupervised learning algorithm, and is an effective method for clustering, visualization and abstraction of complex data (Kohonen, 2001). The main property of the SOM is the dimension reduction achieved by the learning process. Input information of high-dimensional input patterns is mapped into low dimensions (commonly a two-dimensional map), with information on the relationships of input signals preserved as much as possible. Through this process, neighboring input values are mapped onto neighboring (or the same) nodes according to some metric defined in the output space and neighborhood topology is preserved. Additionally, the SOM averages the input dataset in weight vectors through the learning process and thus removes noise (Vesanto et al., 1998). The averaging effect was implemented in the first step of the combined model in this study. It is well known that classical clustering techniques are sensitive to the presence of outliers in the data. Therefore, outliers must be detected before analyzing clusters. However, the problem of outliers is minimized in the SOM. Each outlier takes its place in one unit of the map, and only the weights of that unit and its nearest neighbors are affected. There is no effect on the other units. Likewise, scattered nodes on the map in turn suggest the presence of an outlier (Lek and Guégan, 2000).

Meanwhile, the ART also clusters input data by self-organizing. But the process of self-stabilization is conducted through different pathways. While differences in neighbor nodes are gradually intensified through competitive learning in the SOM, the large clusters were successively divided into subclusters (Fig. 4). The ART takes a set of input vectors and gives, as output, a set of clusters that map each input vector to a cluster. Input vectors are mapped to output vectors according to their closeness, which is determined by a similarity measure. The clusters can then be labeled to indicate their semantic meaning and are represented internally by using prototype vectors. Both the SOM and the ART are used for clustering datasets. However, the main difference concerns the topology preservation of the output units of networks: neighboring topology is preserved in SOM, while it is not preserved in ART. Owing to the topology preservation of the SOM, the SOM is preferred in the ordination of samples rather than the ART. The ART, however, is superior in clustering the groups in different scales according to the vigilance residing in the network (Lin and Lee, 1996). In this study the SOM was used to visualize the topology of neighboring nodes for the first step, and the ART was subsequently used to classify the patterned nodes in the SOM topography in different hierarchical levels.

Through hierarchical classifications, samples can be efficiently assessed through different levels of association. As mentioned before, the upstream (HD1-2) and downstream (HD4-5) samples in Hoedong were grouped differently in cluster III and in cluster I, respectively, in the highest hierarchical level (Figs. 2 and 4). This indicates that the communities in these two areas were very different although they were located close to each other in the same stream (Fig. 1). The degree of organic pollution showed a distinctive difference between these two groups of the sample sites. Sites HD4-5 were heavily polluted, being affected by domestic sewage and waste from a junkyard located near the sampling areas, whereas sites HD1-2 were relatively clean.

Community composition also confirmed the differences between the two groups of the samples. *A. hilgendorfi* and *C. dentriculata* were exclusively collected at sites HD1-2 (Figs. 2 and 6). *Asellus* is an indicator species of β-mesosaprobity and is collected in a moderately enriched region with rich algal growth (Wiederholm, 1984; Kwon and Chon, 1991). Furthermore, macrophytes *Hydrilla verticillata* and *Potamagetum criptus* and filamentous algae *Oedogonium* were abundant at these sample sites. These taxa serve as a major energy source for the macroinvertebrates. At sites HD4-5, in contrast, species richness was low and a few selected species tolerant to pollution were dominant, including *Chironomus* sp. and *L. hoffmeisteri* (Figs. 2 and 6). In this case, domestic organic waste was the main food supply for the tolerant species.

In the SOM, samples in HD3 were mostly located between HD1-2 and HD4-5 at boundary areas (Fig. 2). The field data also confirmed that community compositions were intermediate between HD1-2 and HD4-5. The BOD values also were in the middle range (Fig. 1b). The benthic communities in two streams, Suyong and Cheolma, were patterned differently in the hierarchical classification of the SOM. Although their BOD values were similar (2.35 ppm and 2.06 ppm in Suyong and Cheolma, respectively; *t*-test, $P>0.05$), community compositions were distinctly different between these streams. Kwon and Chon (1991) reported that the taxa were more diverse and a selected taxon such as Chironomidae occurred more abundantly in Suyong than in Cheolma. These differences of community compositions in the two Suyong and Cheolma streams were also clearly addressed in the hierarchical patterning of the SOM: the groups in Suyong and Cheolma were separated across different hierarchical levels (Fig. 4a–c). These results demonstrate that communities could be grouped in different scales, revealing an extra degree of explanation of ecological states. The univariate chemical assessment such as BOD or biological index such as BMWP alone cannot convey information residing in ecosystem data as efficiently as community mapping.

Communities, in general, usually develop into hierarchical organizations, and it is helpful to understand communities at different scales (Allen and Starr, 1982; O'Neill et al., 1986; Urban et al., 1987; Allen and Hoekstra, 1992). The study of complex systems emphasizes the importance of scale (O'Neill, 1989; Levin, 1992), and developments in hierarchy theory demonstrate how processes and constraints change across different scales (Allen and Starr, 1982; O'Neill et al., 1986). Benthic macroinvertebrate communities in streams usually have clear taxonomic and functional hierarchies, and these are essential to verify the organizational characteristics in communities (Cummins et al., 1973; Cummins, 1974). Hierarchical grouping was demonstrated by combinational implementation of artificial neural networks in this study. Although the hierarchical grouping was only conducted on the community composition level, the type of hierarchical model presented in this study could be universally applied to revealing community organization in various domains (e.g. food web, trophic structure, etc.) if appropriate data are available.

Hierarchical classification would be also useful for diagnosing spatial and temporal variations of communities. The present study demonstrated the usefulness tracking ecological states of community development on different hierarchical levels. As shown in Fig. 5, new datasets were validated, and changes in states of communities across different hierarchical levels were demonstrated. In hierarchical grouping, there were nodes where no samples were clustered. In subclusters Ic2 and Id2 (Fig. 4c), for instance, no communities were observed. Although no communities reside there, these clusters are still useful and serve as reference nodes for validation. For example, if data from a new community belong to this node, ecological states could be considered in terms of differences in hierarchical clustering. With hierarchical grouping, disturbance can be diagnosed in spatial and temporal domains on different scales.

To cluster the nodes of the SOM, the U-matrix algorithm is conventionally used: the matrix gives a picture of the topology of the unit-layer and therefore also of the topology of the input space (Ultsch, 1993). Sometimes, however, it is not an easy task to detect clear boundaries on the map of the U-matrix. The feasibility of grouping between the U-matrix and hierarchical classification was compared in this study (Fig. 4). It appeared that grouping was more apparent in hierarchical cluster analysis than in U-matrix, responding to environmental effects (e.g. Figs. 4–6). However, the U-matrix presented effectively overall similarities between the units of the SOM map. This confirmed the work of Vesanto and Alhoniemi (2000), reporting that the U-matrix did not present clusters in their datasets, while an agglomerative clustering method showed clear clusters. In ecological studies, several different cluster algorithms such as a fuzzy c-means clustering method (FCM; Giraudel et al., 2000), a k-means algorithm (Park et al., 2003a) and an agglomerative clustering method (Park et al., 2003b) have been used to cluster the nodes of the trained SOM map. Different methods to group the map have both strengths and weaknesses according to their clustering algorithms (Jongman et al., 1995; Legendre and Legendre, 1998; Vesanto and Alhoniemi, 2000). However, it is not easy to compare efficiency in grouping since the data were basically trained in an unsupervised manner so no reference (i.e. template or teacher) is available for comparison. Further study is required to improve the methods used for quantitatively determining the feasibility of grouping.

Comparing the SOM and the PCA, grouping by the SOM was more relevant to ecology, revealing different effects of pollution states, and impacts of spatial and temporal variations in environment (Figs. 2, 6 and 7). The SOM, by explaining total variance in the data, was able to describe more directly the discriminatory power of input variables in mapping, while PCA explained less than 30% of the total variance in the data. Since the SOM is efficient in non-linear classification, ordination and visualization, the horseshoe effect observed in PCA doest not exist in the SOM. In the SOM map, however, the direction of the gradients cannot be controlled quantitatively as in PCA, since it is based on a heuristic and adaptive learning algorithm. However, the SOM has an additional advantage. It is superior in visualization of variables. For example, species abundance and richness, and environmental factors can be visualized in the same SOM map, efficiently elucidating relationships between variables. In this study, as stated above, the distances between the groups in the SOM map were efficiently defined through hierarchical levels by implementing two-level classifications. It showed the degree of association among the sample communities in a hierarchy (Fig. 4), alleviating the problem of objectivity in distance on the SOM nodes (Chon et al., 1996).

Similarly to correspondence analysis, the SOM can display groupings of samples and species simultaneously. Through this visualization, the importance of each species can be evaluated (Fig. 6) and the affinity between species in the community assemblages can be described. Another advantage of the SOM is that new samples can be added (or tested) on the SOM map without affecting the previously trained ordination. Through this process, we evaluated new samples (collected 1993–1995) in the SOM, and effectively traced seasonal changes in communities (Fig. 5).

In conclusion, the two-level classification approach by combining the SOM and the ART was efficient for a comprehensive understanding of multivariate ecological data. It can be efficiently utilized for diagnosing changes in ecological states across different hierarchical levels in spatial and temporal domains. The combined network of unsupervised learning could be a useful tool for ecosystem managers in the monitoring and assessment of disturbances in ecosystems.

Acknowledgments

This work was supported in part by the Postdoctoral Fellowship Program of Korea Science and Engineering Foundation, grant No. R01-2001-000-00087-0 from Korea Science and Engineering Foundation, and the EU project PAEQANN (No EVK1-CT1999-00026).

References

Aguilera PA, Frenich AG, Torres JA, Castro H, Vidal JLM, Canton M. Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. Water Res 2001;35:4053–4062.

Alhoniemi E, Himberg J, Parhankangas J, Vesanto J. SOM Toolbox. [online] http://www.cis.hut.fi/projects/somtoolbox, 2000.

Allen TFH, Hoekstra TW. Toward a unified ecology. New York: Columbia University Press, 1992. p. 384

Allen TFH, Starr TB. Hierarchy: perspectives for ecological complexity. Chicago: The University of Chicago Press, 1982. p. 310

Baraldi A, Alpaydin E. Simplified ART: a new class of ART algorithms. Berkeley:International Computer Science Institute, TR-98-004, 1998; p. 42.

Barbour MT, Gerritsen J, Snyder BD, Stribling JB. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish, Second Edition. EPA 841-B-99-002. US Environmental Protection Agency; Office of Water; Washington, DC, 1999; p. 339.

Beals EW. Bray–Curtis ordination: an effective strategy for analysis of multivariate ecological data. Adv Ecol Res 1984;14:1–55.

Carpenter GA, Grossberg S. ART2: self-organization of stable category recognition codes for analog input patterns. Appl Optics 1987;26(3):4919–4930.

Céréghino R, Giraudel JL, Compin A. Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self organizing maps. Ecol Model 2001;146:167–180.

Chon T-S, Park YS, Moon MH, Cha EY. Patternizing communities by using an artificial neural network. Ecol Model 1996;90:69–78.

Chon T-S, Park Y-S, Park J-H. Determining temporal pattern of community dynamics by using unsupervised learning algorithms. Ecol Model 2000;132:151–166.

Coysh J, Nichols S, Ransom G, Simpson J, Norris R, Barmuta L, Chessman B. AUSRIVAS Predictive Modelling Manual. http://ausrivas.canberra.edu.au/ 2000.

Cummins KW, Petersen RC, Howard FO, Wuycheck JC, Holt VI. The utilization of leaf litter by stream detritivores. Ecology 1973;54:336–345.

Cummins KW. Structure and function of stream ecosystems. Bioscience 1974;24:631–641.

Giraudel JL, Aurelle D, Berrebi P, Lek S. Application of the self-organising mapping and fuzzy clustering to microsatellite data: how to detect genetic structure in brown trout (*Salmo trutta*) populations. In: Lek S, Guégan J-F, editors. Artificial neuronal networks: application to ecology and evolution. Berlin: Springer-Verlag, 2000. p. 187–202.

Giraudel JL, Lek S. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecol Model 2001;146:329–339.

Goodall DW. Objective methods for the classification of vegetation. Aust J Bot 1954;2:304–324.

Hawkes HA. Invertebrates as indicators of river water quality. In: James A, Evision L, editors. Biological indicators of water quality, vol. 2. Chichester, Great Britain: John Wiley and Sons, 1979;2:1–45.

Hawkes HA. Origin and development of the biological monitoring working party score system. Water Res 1997;32:964–968.

Hellawell JM. Biological indicators of freshwater pollution and environmental management. London: Elevier, 1986. p. 546

Hill MO, Gauch HG. Detrended correspondence analysis, an improved ordination technique. Vegetatio 1980;42:47–58.

Hynes HBN. The Biology of polluted waters. London: Liverpool University Press, 1960. p. 202.

Jongman RHG, ter Braak CJF, van Tongerenm OFR, editors. Data analysis in community and landscape ecology. Cambridge: Cambridge University Press, 1995. p. 299.

Kenkel NC, Orloci L. Applying metric and non-metric multidimensional scaling to ecological studies: some new results. Ecology 1986;67:919–928.

Kiviluoto K. Topology preservation in self-organizing maps. In: Pro. ICNN'96, IEE International Conference on Neural Networks IEEE Service Center, Piscataway, 1996; pp. 294–299.

Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43:59–69.

Kohonen T. Self-organizing maps, 3rd ed. Berlin: Springer, 2001. p. 501

Kwon T-S, Chon T-S. Ecological studies on benthic macroinvertebrates in the Suyong River. II. Investigations on distribution and abundance in its main stream and four tributaries. Korea J Lim 1991;24:179–198.

Kwon T-S, Chon T-S. Ecological studies on benthic macroinvertebrates in the Suyong River. III. Water quality estimations using chemical and biological indices. Korea J Lim 1993;26:105–128.

Legendre P, Legendre L. Numerical ecology. Amsterdam: Elsevier, 1998. p. 853.

Lek S, Guégan J-F, editors. Artificial neuronal networks: application to ecology and evolution. Berlin: Springer, 2000. p. 262.

Levin SA. The problem of pattern and scale in ecology. Ecology 1992;73:1943–1967.

Lin CT, Lee CSG. Neural fuzzy systems: a neuro-fuzzy synergism to intelligent systems. Upper Saddle River: Prentice Hall, 1996. p. 797.

Ludwig JA, Reynolds JF. Statistical ecology: a primer on methods and computing. New York: John Wiley and Sons, 1988. p. 337.

McCune, B, Grace JB, Urban DL. Analysis of ecological communities, MjM Software Design, Gleneden Beach, Oregon, USA, 2002; p. 304.

Norris RH. Biological monitoring: the dilemma of data analysis. J N Am Benthol Soc 1995;14:440–450.

O'Neill RV, DeAngelis DL, Allen TFH. A hierarchical concept of ecosystems. Princeton: Princeton University Press, 1986. p. 254.

O'Neill RV. Perspectives in hierarchy and scale. In: May RM, Levin SA, editors. Perspectives in ecological theory-Princeton University Press, 1989. p. 140–156.

Obach M, Wagner R, Werner H, Schmidt H-H. Modelling population dynamics of aquatic insects with artificial neural networks. Ecol Model 2001;146:207–217.

Pao Y-H. Adaptive pattern recognition and neural networks. Reading, Massachusetts: Addison-Wesley, 1989. p. 309.

Park Y-S, Chon T-S, Kwak IS, Kim J-K, Jørgensen SE. Implementation of artificial neural networks in patterning and prediction of exergy in response to temporal dynamics of benthic macroinvertebrate communities in streams. Ecol Model 2001;146:143–157.

Park Y-S, Céréghino R, Compin A, Lek S. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol Model 2003a;160:265–280.

Park Y-S, Chang J, Lek S, Cao W, Brosse S. Conservation strategies for endemic fish species threatened by the Three Gorges Dam. Conserv Biol 2003b;17:1748–1758.

Pearson K. On lines and planes of closest fit to systems of points in space. Philos Mag 1901;2:559–572.

Recknagel F, editor. Ecological informatics: understanding ecology by biologically-inspired computation. Berlin: Springer, 2002. p. 398.

Spellerberg IF. Monitoring ecological change. Cambridge: Cambridge University Press, 1991. p. 334.

Statsoft, Inc. STATISTICA (data analysis software system), www.statsoft.com. 2000.

The Mathworks, lnc. MATLAB Version 6.1, Massachusetts, 2001.

Ultsch A. Self-organizing neural networks for visualization and classification. In: Opitz O, Lausen B, Klar R, editors. Information and classification. Berlin: Springer-Verlag, 1993. p. 307–313.

Urban DL, O'Neill RV, Shugart HH. Landscape ecology: a hierarchical perspective can help scientists understand spatial patterns. Bioscience 1987;37:119–127.

Vesanto J, Alhoniemi E. Clustering of the self-organizing map. IEEE T Neural Network 2000;11:586–600.

Vesanto J, Himberg J, Alhoniemi E, Parhankangas J. Self-organizing map in Matlab: the SOM Toolbox. In Proceedings of the MATLAB Digital Signal Processing Conference, Espoo, Finland, 1999; pp. 35–40.

Vesanto J, Himberg J, Siponen M, Simula O. Enhancing SOM based Visualization. In Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems (IIZUKA'98), Iizuka, Japan, 1998; pp. 64–67.

Walley WJ, Hawkes HA. A computer-based reappraisal of Biological Monitoring Working Party scores using data from the 1990 River Quality Survey of England and Wales. Water Res 1996;30:2086–2094.

Walley WJ, Hawkes HA. A computer-based development of the Biological Monitoring Working Party score system incorporating abundance rating, biotope type and indicator value. Water Res 1997;31:201–210.

Walley WJ, Martin RW, O'Connor MA. Self-organising maps for the classification of river quality from biological and environmental data. In: Denzer R, Swayne DA, Purvis M, Schimak G, editors. Environmental software systems: environmental information and decision support, IFIP Conference Series, Boston:Kluwer Academic Publishers, 2000; pp. 27–41.

Wiederholm T. Responses of aquatic insects to environmental pollution. In: Resh VH, Rosenberg DM, editors. The ecology of aquatic insects. New York: Praeger Publishers, 1984. p. 508–557.

Wright JF, Furse MT, Armitage PD. RIVPACS: A technique for evaluating the biological quality of rivers in the UK. Eur Water Pollut Control 1993;3:15–25.