available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/ecolinf

Application of a self-organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across France

Young-Seuk Park[a,b,*], *Juliette Tison*[b], *Sovan Lek*[c], *Jean-Luc Giraudel*[d], *Michel Coste*[b], *François Delmas*[b]

[a]*Department of Biology, Kyung Hee University, Hoegi-dong, Dongdaemun-gu, Seoul 130-701, Korea*
[b]*U.R. REQUE, Cemagref Bordeaux, 50 av. de Verdun, 33612 Cestas, France*
[c]*LADYBIO, CNRS-Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse cedex, France*
[d]*EPCA- LPTC, UMR 5472 CNRS-Université Bordeaux 1, 39 rue Paul Mazy, 24019 Périgueux Cedex, France*

ARTICLE INFO

Article history:
Received 9 August 2005
Received in revised form
1 March 2006
Accepted 15 March 2006

Keywords:
Dimension reduction
Representative species
Self-organizing map
Multivariate analysis

ABSTRACT

Ecological communities consist of a large number of species. Most species are rare or have low abundance, and only a few are abundant and/or frequent. In quantitative community analysis, abundant species are commonly used to interpret patterns of habitat disturbance or ecosystem degradation. Rare species cause many difficulties in quantitative analysis by introducing noises and bulking datasets, which is worsened by the fact that large datasets suffer from difficulties of data handling. In this study we propose a method to reduce the size of large datasets by selecting the most ecologically representative species using a self organizing map (SOM) and structuring index (SI). As an example, we used diatom community data sampled at 836 sites with 941 species throughout the French hydrosystem. Out of the 941 species, 353 were selected. The selected dataset was effectively classified according to the similarities of community assemblages in the SOM map. Compared to the SOM map generated with the original dataset, the community pattern gave a very similar representation of ecological conditions of the sampling sites, displaying clear gradients of environmental factors between different clusters. Our results showed that this computational technique can be applied to preprocessing data in multivariate analysis. It could be useful for ecosystem assessment and management, helping to reduce both the list of species for identification and the size of datasets to be processed for diagnosing the ecological status of water courses.

© 2006 Elsevier B.V. All rights reserved.

1. Introduction

Biological communities are commonly used as indicators of ecosystem quality. Community structures are determined by many environmental factors in different spatial and temporal scales (Stevenson, 1997; Snyder et al., 2002). Community data are composed of a large number of species collected at many sampling sites at different times. A commonly observed phenomenon in field surveys is that the vast majority of species are represented by low abundance while only a few species are abundant. Preston's canonical log-normal distribution is the most widely accepted formalization of the relative commonness and rarity of species (Preston, 1962; Brown, 1981).

* *Corresponding author*. Department of Biology, Kyung Hee University, Hoegi-dong, Dongdaemun-gu, Seoul 130-701, Korea. Tel.: +82 2 961 0946; fax: +82 2 961 0244.
E-mail address: parkys@khu.ac.kr (Y.-S. Park).

1574-9541/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2006.03.005

In quantitative community analysis, abundant species are commonly used to interpret patterns of habitat disturbance or ecosystem degradation, whereas rare species are generally excluded from the analysis. Although the effects of rare species are negligible on statistical results, they introduce noise and cause difficulties in data analyses. By removing noise, the more important information is more likely to be detected (McCune et al., 2002). To solve the problems of rare species in community ecology, several different approaches (i.e., down weighting, overweighting and deleting species) are applied depending on researchers' interests (Mante et al., 1995, 1997; Cao et al., 2001; Fodor and Kamath, 2002). This is regarded as a preprocessing stage in data mining. As illustrated in Fig. 1, data mining consists of two main steps, data preprocessing and pattern recognition (Fodor and Kamath, 2002). Preprocessing is often time consuming, yet critical as a first step. To ensure the success of the data mining process, it is important that the features extracted from the data should be representative of the data to be relevant to the issues for which the data are collected.

In community ecology, ordination and classification techniques are commonly used to simplify the interpretation of a complex dataset. However, this purpose is defeated if there are a very large number of variables. A large number of variables in the analysis may be informative to investigators in the exploratory phase of the study, yet it is difficult to point out the major issues contained in the dataset if the ordination diagrams are cluttered by numerous variables (Palmer, 2005). Therefore, it is desirable to reduce the number of variables for multivariate analysis in many cases. However, it is impossible to reduce the number of variables without the risk of losing information. In order to remove variables, one should make sure that ecologically relevant information is retained as far as possible.

Deleting rare species could be a useful way of reducing the bulk of ecological datasets and noise generated without losing much information (McCune et al., 2002). The simplest way to delete rare species is to consider the frequency of species in samples (MJM Software Design, 2000), and to carry out direct or indirect gradient analyses including Principal Component Analysis, Correspondence Analysis, Detrended Correspondence Analysis, Canonical Correspondence Analysis, etc. However, traditional multivariate analyses are generally based on linear principles (James and McCulloch, 1990), and cannot overcome various problems: biases due to complexity and non-linearity residing in datasets, and inherent correlations among variables (Lek et al., 1996; Brosse et al., 1999). Self-organizing map (SOM) (Kohonen, 1982), on the other hand, has been used as an alternative to traditional statistical methods to efficiently deal with datasets ruled by complex, non-linear relationships (Lek et al., 1996; Lek and Guégan, 2000). The SOM, an unsupervised neural network, has been implemented to analyse various ecological data (Lek and Guégan, 1999, 2000; Recknagel, 2003): evaluation of environmental variables (Park et al., 2003a; Céréghino et al., 2003), classification of communities (Chon et al., 1996; Park et al., 2003b; Tison et al., 2005), water quality assessments (Walley et al., 2000), and prediction of population and communities (Céréghino et al., 2001; Obach et al., 2001). The SOM produces virtual communities in a low dimensional lattice through an unsupervised learning process. Input components (i.e., species) could be visualized on a SOM map to show the contribution of each component in the self-organization of the map (Park et al., 2003b). These component planes can be considered as a sliced version of the SOM map and provide a powerful tool to analyze the community structure. But, when we consider a lot of species (i.e., several hundreds or thousands), it is difficult to compare all component planes for all species. It becomes necessary to develop an efficient method to select species for removal.

In this study we propose a computational method to reduce the number of species in datasets with a large number of

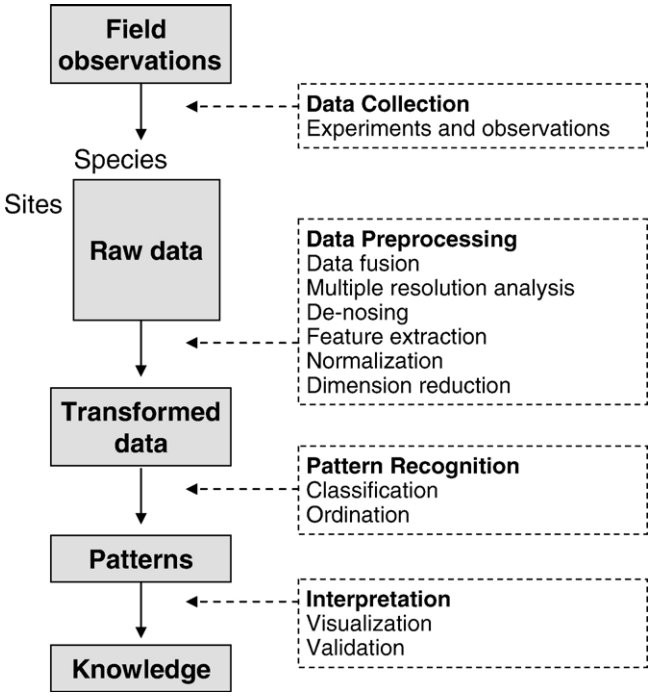

Fig. 1 – Schematic diagram of a data mining process.

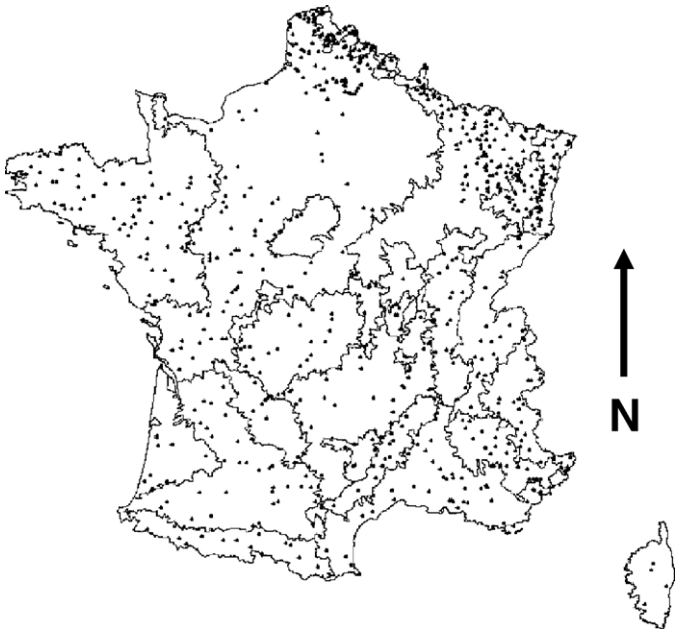

Fig. 2 – Distribution of diatom sampling sites in a French hydrosystem.

species without losing much information. The datasets with the reduced number of species were further evaluated in relation to environmental conditions. This approach can contribute to practical ecosystem management in handling huge datasets and would broaden the scope of SOM in mining community data in diverse quantitative ecological studies.

2. Materials and methods

2.1. *Ecological dataset*

From the Cemagref French Diatom Database, 836 samples were extracted. The data had been collected nationwide throughout France (Fig. 2) in summer from 1979 to 2002 according to the NFT 90-354 recommendations (AFNOR, 2000). Diatom species were identified at a 1000× magnification (Leitz DMRD photomicroscope) according to Krammer and Lange-Bertalot (1986, 1988, 1991a, 1991b): examination of permanent slides of cleaned diatom frustules, having been digested in boiling H_2O_2 (30%) and HCl (35%), and mounted in a high refractive index medium (Naphrax, Northern Biological Supplies Ltd, UK; RI=1.74). A relative abundance of species was obtained by randomly selecting 400 individuals per sample for taxonomic identification to species level.

Among the 941 species recorded in the dataset, 490 were observed in less than 10 samples (Fig. 3). More than 52% of species were only identified in less than 1.2% of samples. Some rare species, which are ecologically important, showed middle or high abundance but occurred only in a limited number of samples. They characterize particular types of environmental conditions, for example *Eunotia exigua* for acidic rivers. Such species must be considered as important, if we want to extract the most relevant ecological information from the datasets although their occurrence numbers are low. On the other hand, about 3% abundant species (25 species) were observed in more than 50% of samples. In particular, the species *Achnanthidium minutissimum* was most frequently observed in 737 samples. A few intermediately tolerant species are also wide spread in the dataset, like *Navicula cryptotenelloides*. Overall, a large variation in abundance was observed in the dataset.

The original dataset consisted of 836 samples with 941 species. The species abundance was transformed by natural logarithm. To avoid a problem of logarithm zeros, the number 1 was added to the density of each species. Subsequently the transformed data were proportionally scaled between 0 and 1 over the range of the minimum and maximum abundance for each species. Through these procedures, the weights (i.e., importance) for the species with low abundance were accordingly increased.

2.2. *Overall modelling procedure*

With the rescaled dataset, SOM classified samples in 2D space and produced weight vectors representing the approximation of

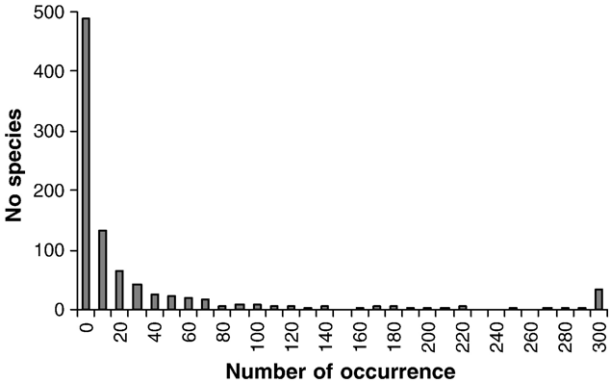

Fig. 3 – Distribution of occurrence frequency of diatom species in the dataset.

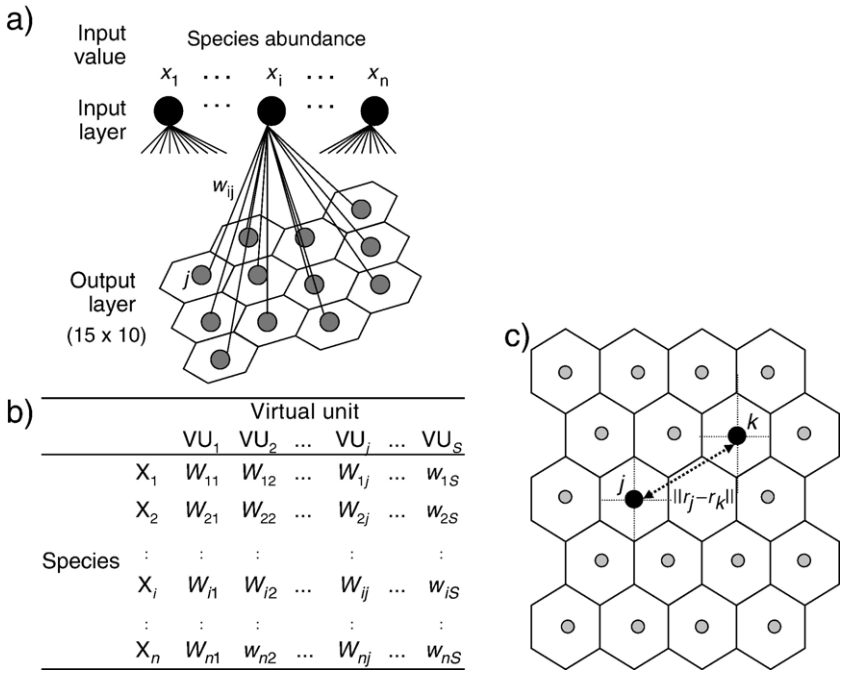

Fig. 4 – Schematic diagram of SOM (a), data structure of virtual community units produced in the SOM learning process (b), and topological distance of the SOM output units used in the SI calculation (c).

input data and typical community types. To quantify the contribution of each species in SOM patterning, a structuring index (SI) (Park et al., 2005) was calculated using prototype vectors of SOM. Subsequently, several different datasets were produced based on the SI histogram by deleting species with low SI in each class of the histogram. These new datasets were trained separately with a new SOM. New SI values were calculated for each species in different datasets. Finally, we computed squared Euclidean distances of SI between the original dataset and reduced datasets. Based on the distances, we choose a criterion for the species to be selected for removal from the datasets while minimizing the loss of ecological information.

2.3. *Self-organizing map (SOM)*

The SOM approximates the probability density function of input data through an unsupervised learning algorithm, and is an effective method for clustering, but also for the visualization and abstraction of complex data (Kohonen, 2001). The algorithm has properties of neighborhood preservation and local resolution of the input space proportional to the data distribution (Kohonen, 1982, 2001). The SOM is widely applicable to the fields of data management, such as data mining, classification, and biological modelling in terms of a nonlinear projection of multivariate data into lower dimensions (Lek and Guégan, 2000; Kohonen, 2001; Park et al., 2003a, 2003b). The SOM consists of two layers: an input layer formed by a set of nodes (or neurons which are computational units), and an output layer formed by nodes arranged in a two-dimensional grid (Fig. 4a). In this study, each input node accounts for the abundance of each species. The output layer was made of a total of S output nodes in the hexagonal lattice (i.e., 150 nodes in a grid of 15×10 cells in this study) for providing better visualization. A hexagonal lattice is preferred because it does not favor horizontal or vertical directions (Kohonen, 2001). The number of nodes was determined as $5x\sqrt{\text{number of samples}}$ (Vesanto, 2000). Subsequently the map size was determined. Basically, the two largest eigen values of the training data were calculated and the ratio between side lengths of the map grid was set to the ratio between the two maximum eigen values. The actual side lengths were then set so that their product was close to the determined number of map units as stated before.

In this study, each sample has been assigned to one output node as a result of SOM calculation. Each output node has a vector of coefficients associated with input data. The coefficient vector is referred to a weight (or connection intensity) vector W between input and output layers. The weights establish a link between the input units (i.e., species) and their associated output units (i.e., groups of samples). Therefore, the output units are referred to virtual community units representing typical community composition of samples assigned in the output units (Fig. 4b). Each vector of each virtual community unit is referred to a prototype vector.

The algorithm can be described as follows: when an input vector X (in this case, the relative abundance of 941 species in a sample) is presented to the SOM, the nodes in the output layer compete with each other, and the winner (whose weight is the minimum distance from the input vector) is chosen. The winner and its neighbors predefined in the algorithm update their weight vectors according to the SOM learning rules as follows:

$$w_{ij}(t+1) = w_{ij} + \alpha(t) \cdot h_{jc}(t)[x_i(t) - w_{ij}(t)] \quad (1)$$

where $w_{ij}(t)$ is a weight between a node i in the input layer and a node j in the output layer at iteration time t, $\alpha(t)$ is a learning

rate factor which is a decreasing function of the iteration time t, and $h_{jc}(t)$ is a neighborhood function (a smoothing kernel defined over the lattice points) that defines the size of neighborhood of the winning node (c) to be updated during the learning process. This learning process is continued until a stopping criterion is met, usually, when weight vectors stabilize or when a number of iterations are completed. This learning process results in the preservation of the connection intensities in the weight vectors.

2.4. *Structuring index (SI)*

The SI was originally developed to define species showing the strongest influence on the organization of the SOM map (Park et al., 2005). Tison et al. (2004, 2005) used the SI to evaluate relevant diatom species in the classification of diatom communities. The SI is the value indicating the relative importance of each species in determining the distribution patterns of the samples in the SOM. Therefore, the set of species showing high SI can be considered as the indicator species.

The SI is calculated from the sum of the ratios of the distance between the weights (i.e., connection intensities) of all species in the SOM and the topological distance between two SOM units (Fig. 3c). This results in representing distribution gradients for each species in the trained SOM. A structuring index of species i, SI_i, is expressed in the equation as follows:

$$SI_i = \sum_{j=1}^{S} \sum_{k=1}^{j-1} \frac{|w_{ij} - w_{ik}|}{||r_j - r_k||} \quad (2)$$

where w_{ij} and w_{ik} are respectively the connection weights of species i (in the input layer) in SOM units j and k, $\|r_j - r_k\|$ is the topological distance between units j and k, and S is the total number of SOM output units. SI considers the distribution gradients of each species in the SOM map. Species showing a strong gradient display a high SI value, whereas species showing a weak gradient present a low SI value. Thus, the higher the value of SI, the more relevant the variable is to the structure of the map.

3. Results

3.1. *Patterning samples with a large dataset*

Diatom communities consisting of 941 species were patterned through the learning process of the SOM (Fig. 5a). Grey scale

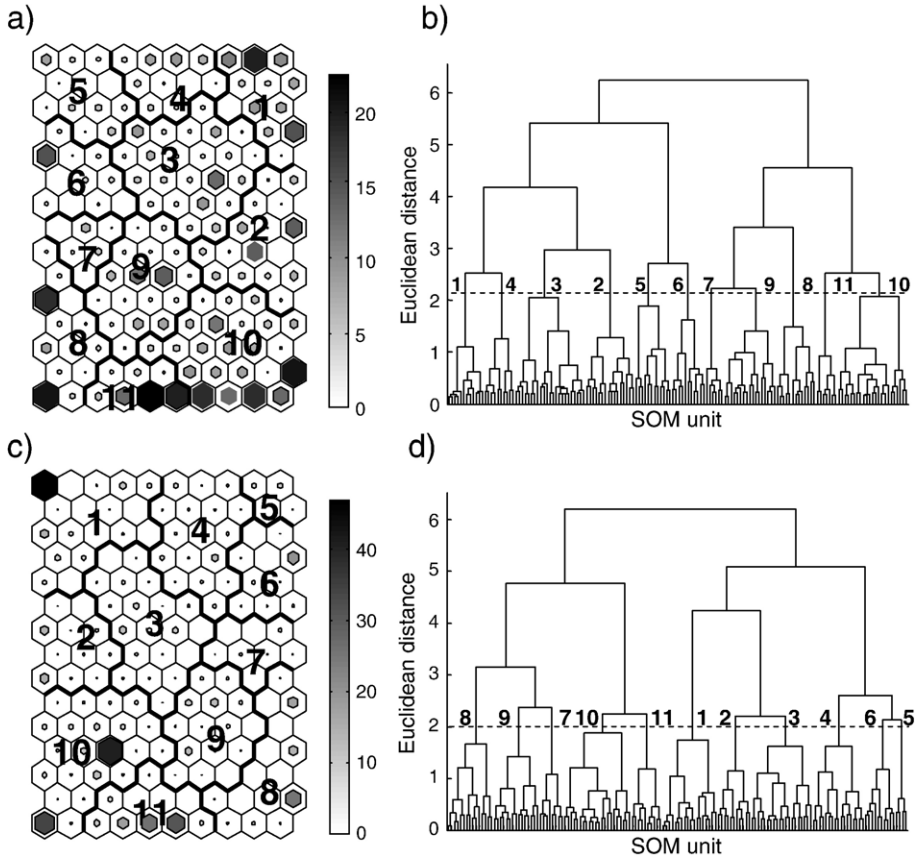

Fig. 5 – Classification of 836 samples through the training of SOM with 941 species (a, b) and 353 species (c, d). Gray scale hexagons in each SOM unit represent the number of samples assigned to each SOM unit in the range of scale bars. Sample names were not given in the SOM units because of limited space. The SOM units were classified into 11 clusters based on the dendrogram of the hierarchical cluster analysis using Ward's linkage method with the Euclidean distance measure (b; for 941 species dataset, and d; for 353 species dataset). The smallest branches in the dendrogram represent SOM units. The unit numbers were not presented due to the small space.

hexagons represent the number of samples assigned in each SOM unit in the range of 2 (small white)–22 (large black). The SOM units were further grouped into 11 clusters based on the dendrogram of a hierarchical cluster analysis (Fig. 5b).

The SOM weight vectors were used for the classification of the units. Overall diatom communities were well organized in the SOM map according to similarities of their species composition. Each cluster was characterised by the ecological conditions and pollution levels of the samples (Fig. 6a). The variation of each environmental parameter was represented with a 95% confidence interval. All 8 environmental variables were significantly different between clusters (Kruskal–Wallis test, $P<0.001$).

Through the SOM learning process, the weight vector was approximately proportional to the probability density of the

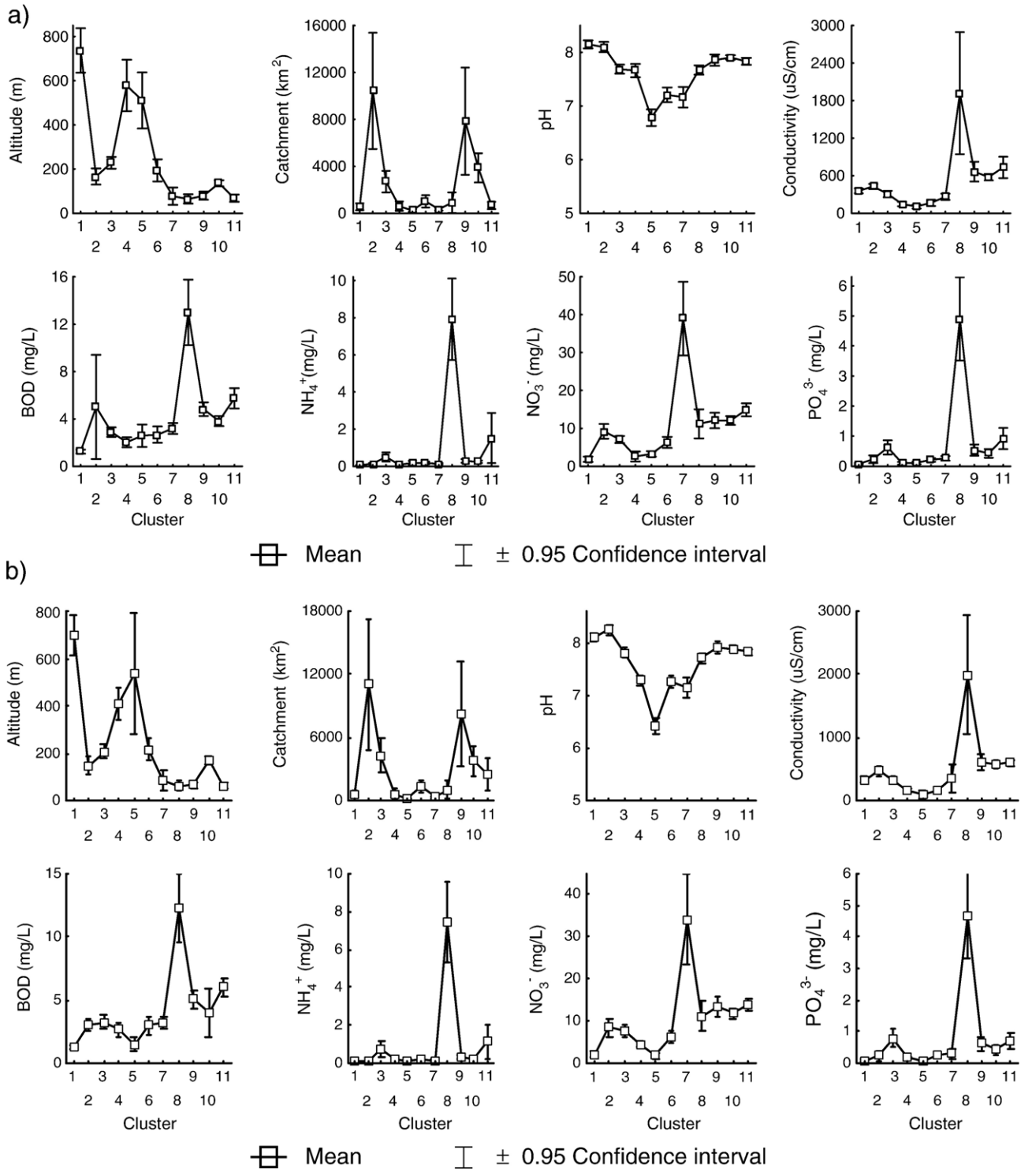

Fig. 6 – Differences of 8 environmental factors at different clusters defined in the SOM map trained with 941 species (a) and 353 species (b).

data. Therefore, each species distribution in the SOM output units can provide their importance in the community structure. Fig. 7a shows examples of distribution gradients of species in the SOM map trained with 941 species. Dark represents a high value of species density in their given scale bar, whereas white is a low value. The values indicate estimated abundances of species in log scale which were denormalized from weight vectors based on the minimum and maximum values of each species defined in the input dataset. All species showed the strong gradient in different ways, although some species showed bi- or multi-modal distribution patterns. While some species showed very similar patterns of gradient on the map, their contributions to patterning on the map were different by displaying different abundances and SI. For instance, *Eunotia bilunaris* and *Achnanthidium eutrophilum* were mainly distributed in the samples assigned to the upper left areas of the SOM map. However, estimated abundances between two species were strongly variable, indicating differences of their contributions to community patterns. The same situation was observed between *Achnanthidium subatomus* and *Surirella linearis* and between *Skeletonema potamos* and *Gomphonema entolejum*. From these types of visualizing component planes, we can evaluate the relative importance of each species. For instance, *A. subatomus* is more important in characterizing samples belonging to the middle upper areas of the SOM map than *S. linearis*.

However, evaluation of contributions becomes difficult from component planes when the numbers of input variables (i.e., species) are very large as stated before. In this study the relative importance of each species was expressed through the SI. The priority of selection for the datasets was determined by the values of the SI (Fig. 8). The SI values of example species are given in Fig. 7, while a profile of the SI

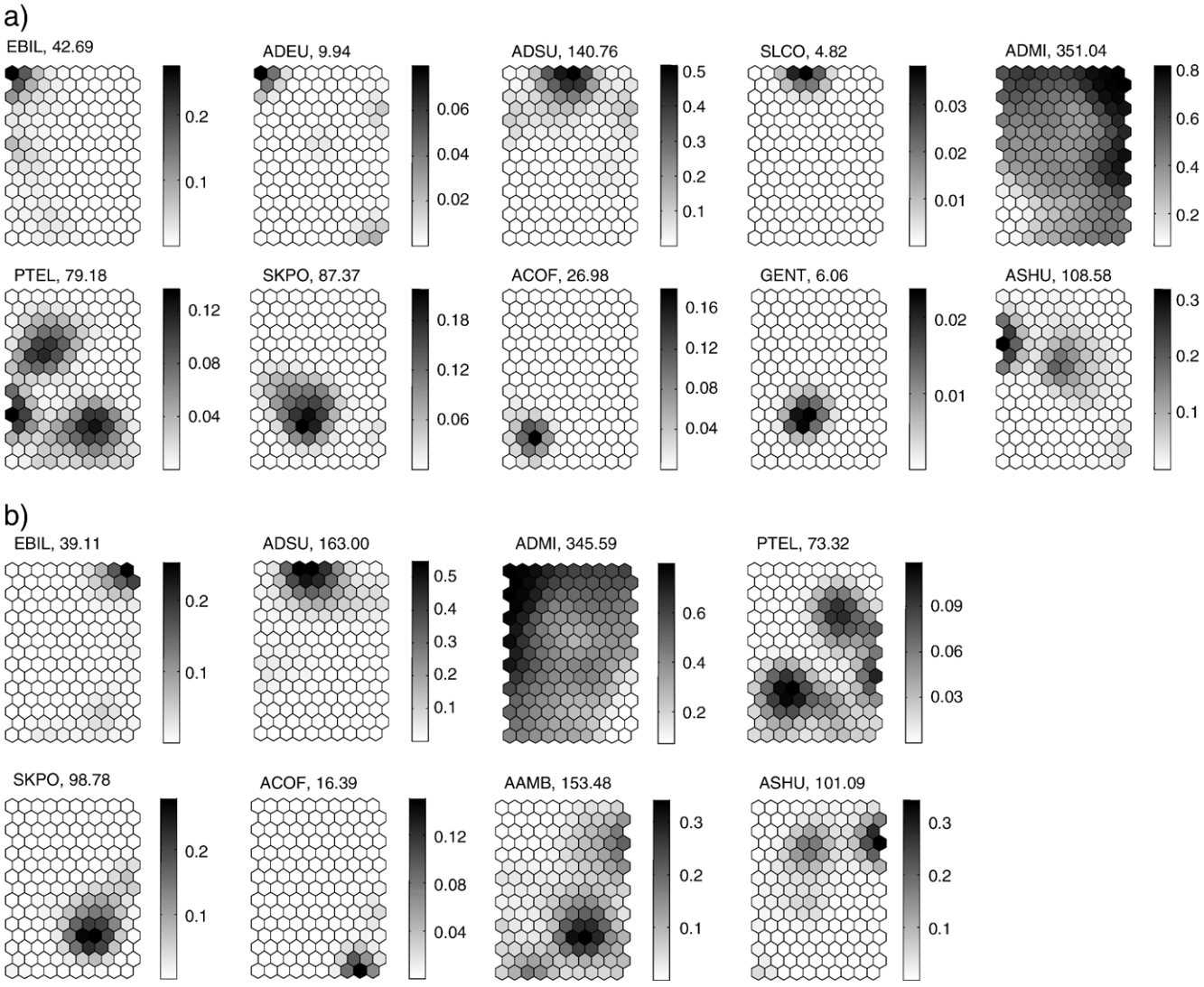

Fig. 7 – Gradient distributions of example species in the SOM map trained with 941 species (a) and 353 species (b). Values following species acronyms are the SI values for each species. Scale bar shows species abundance calculated through SOM learning process in log scale. AAMB, Aulacoseira ambigua, ACOF; Amphora coffeaeformis, ADEU; A. eutrophilum, ADMI; A. minutissimum, ADSU; A. subatomus, ASHU; Achnanthes subhudsonis, EBIL; E. bilunaris, GENT; G. entolejum, PTEL; Planothidium ellipticum, SKPO; S. potamos, SLCO; S. linearis.

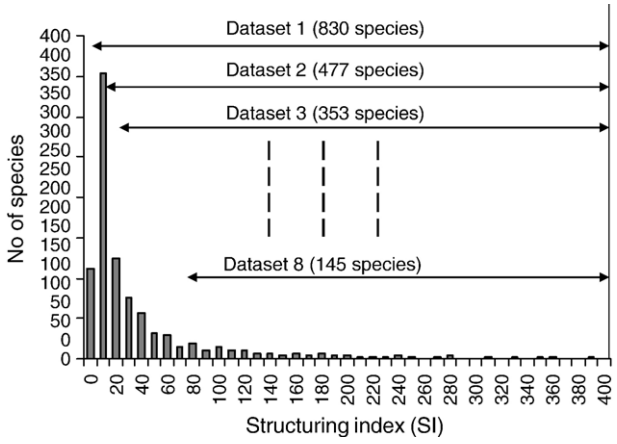

Fig. 8 – Number of species at different classes of SI in the original dataset containing 941 species. Based on the SI classes, 8 different datasets were built by excluding species showing low SI.

values over 941 species is given in different classes in Fig. 8. More than 49% of species showed less than 20 SI values (in the first and the second classes from the left on the *x* axis). The number of species decreased gradually until the 8th class (60–70 SI). The contributions of most species beyond the 8th class were very low in defining community patterns.

3.2. *Selection of relevant species*

The next step is to evaluate different datasets where relevant species were selected according to values of SI. SOM trainings were independently repeated with 8 successive datasets as shown in Fig. 8. The SI values were summed for all species in each dataset. Subsequently the sum of Euclidean distances of the SI values between the original dataset and the reduced datasets were calculated (Fig. 9). Datasets consisting of a small number of abundant species showed high SI values for most species, whereas larger datasets with diverse species covered species with both high and low SI values. As the number of species decreased, the Euclidean distance increased abruptly around the number of species approximately equal to 353 whereas the distances were gradually decreased as the number of species further increased to 941. The profile of the distances indicates that elimination of species after 353 would not seriously affect the predictive characteristics of the original dataset. Consequently the profile of the distances (Fig. 9) shows a criterion to choose the SI value to select the appropriate number of species while minimizing the loss of information due to removal of extra species. Here, we chose the 3rd class with 353 species based on the above reasoning. Regression analyses were carried out between SI values of the total species (941) and each of the datasets with the reduced number of species. In the case of the 353 species dataset, the regression determination coefficient was distinctively high (R^2=0.991) (Fig. 10) while the coefficients for the datasets with a lower number of species (145 and 220) were substantially lower. This indicates that the community information is preserved in a new reduced dataset with 353 species.

3.3. *Patterning samples with reduced dataset*

To evaluate the dataset with 353 relevant species, 836 samples were trained with the SOM (Fig. 5c). The numbers of samples assigned to each SOM unit are indicated in a grey scale as hexagons ranging from 0 (small white) to 47 (large black). The SOM units were classified into 11 clusters through a hierarchical cluster analysis based on Ward's linkage method (Fig. 5d). The grouping was essentially the same as those of the original dataset (Fig. 5a,b). Overall, diatom communities were well organized in the SOM map according to similarities of species composition. Each cluster was well characterised by the ecological and pollution conditions of the sampling sites (Fig. 6b). All 8 environmental variables were significantly different between clusters (Kruskal–Wallis test, $P<0.001$). Samples in clusters 1–6 (in the upper areas of the SOM map) were mainly collected from the sites showing higher water quality, whereas the samples in clusters 7–11 (in the bottom areas of the SOM map) were observable from disturbed sites. The difference of communities (communities in good water quality *versus* communities under anthropogenic disturbances) was clearly distinguished in two main clusters in the dendrogram of the SOM units (Fig. 5d). The clusters were well in accordance with those of the original dataset with 941 species (Fig. 5). The characteristics of each cluster are summarized in Table 1.

Fig. 7b shows the abundance patterns of some selected species in the reduced dataset with 353 species. The patterns were similar to the original dataset (Fig. 7a), although their relative positions in the SOM map were changed to somewhat like mirror images. For example, *E. bilunaris* showed high values in the upper right areas of the SOM map in the original dataset, but was abundant in the upper left areas in the reduced dataset. *A. minutissimum* showed high values in the upper right areas in the original dataset, while the abundance was higher in the upper left areas of the reduced dataset. The results in Figs. 6 and 7 indicate that removal of a substantial number of species according to the SI did not affect the preservation of useful information residing in the original dataset. Species richness between the original dataset and the

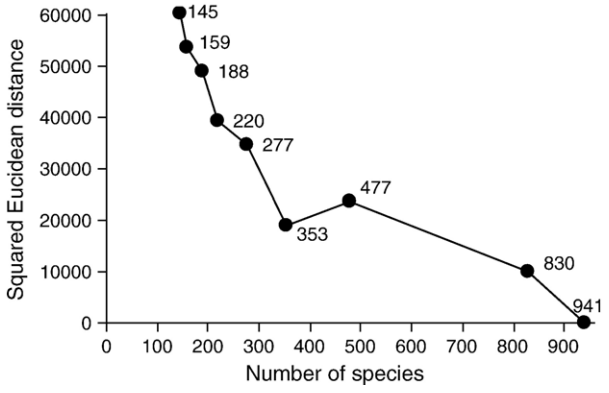

Fig. 9 – Similarity distances of species SI between the original dataset and 8 reduced datasets. In the distance calculations, 145 species included in the smallest dataset were used.

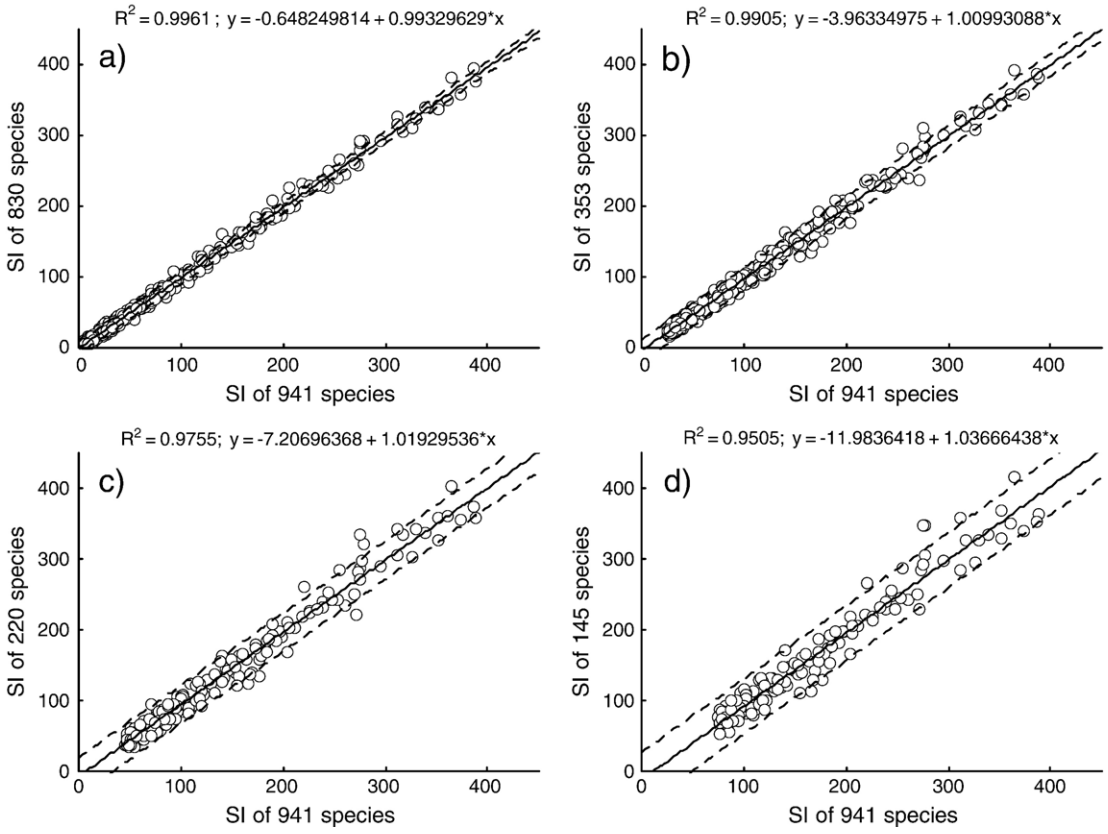

Fig. 10 – Relations of species structuring index (SI) between the original dataset and reduced datasets. The reduced datasets were built by excluding species showing low SI values. a) Dataset of 830 species, b) dataset of 353 species, c) dataset of 220 species, and d) dataset of 145 species. Solid lines represent linear regression and dotted lines are predictive bands giving information on individual predictions of the dependent variable in ±0.95 confidence interval.

reduced dataset with 353 species also showed a strong linear relationship (R^2=0.993).

4. Discussion

Dimension reduction is required when the data are of a higher dimension than tolerated through long-term or large-scale field survey. The goal of dimension reduction is to find a simplified representation of original data without losing much information. Dimension reduction can be considered in two categories: 1) reduction of the number of features representing a data item (from m items in the original data to n items in the reduced data, $n<m$) and 2) reduction of the number of basis vectors used to describe the data (Fodor and Kamath, 2002). In this study we focused on the first category, reduction of the number of features (species) by excluding rare species that would generally induce noise in data. The curse of dimensionality (Bellman, 1961) refers to the fact that, in the absence of simplifying assumptions, the sample size needed to estimate a function of several variables to a given degree of accuracy (i.e., to get a reasonably low-variance estimate) increases exponentially with the number of variables.

A way to avoid the curse of dimensionality is to reduce the input dimensions of the function to be estimated; this is the basis for the use of local objective functions, depending on a small number of variables, in unsupervised methods (Carreira-Perpinan, 2001).

Generally, biological community data consists of many species in many sampling sites for fulfilling various purposes for ecosystem management and installation of management policy for ecosystem health. Diatoms used in this study are also one of the major aquatic taxa considered for monitoring water quality and for aquatic ecosystem management. In addition to the fact that the European Water Framework Directive (European Parliament, 2000) considers benthic diatoms as one of the key organism groups for assessing the ecological quality of rivers, diatom communities are considered as important indicators and have been extensively investigated. However, sampling diatoms gives a large number of species in many samples. It is difficult to manage a huge number of species. A lot of research resources are consumed in identifying species and handling the subsequent datasets. Therefore, it is important to reduce the size of datasets without losing relevant information in diatom data.

In this study, we presented a method to choose relevant species from a large dataset through the learning process of the SOM. Calculation of the structuring power of each species was revealed through self organization. The species selected here (353) showed very similar SOM distribution patterns to

Table 1 – Environmental conditions and representative species in each cluster using 353 selected species

Clusters	Water courses	Environmental conditions	Representative species
1	Upstream	High pH, high altitude	*Achnanthidium biasolettianum, Diatoma ehrenbergii*
2	Downstream	High pH	*Amphora pediculus, Encyonopsis microcephala*
3	Mid-stream	Slightly polluted	*Fistulifera saprophila, Mayamaea atomus*
4	Upstream	Moderately polluted, high altitude	*Achnanthidium subatomus, Fragilaria arcus*
5	Upstream	Low pH, low conductivity	*Eunotia exigua, Frustulia saxonica*
6	Downstream	Low conductivity	*Navicula rhynchocephala, Gomphonema exilissimum*
7	Downstream	High NO_3^-	*Navicula gregaria, Eolimna minima*
8	Downstream	Heavily polluted, low altitude, high PO_4^{3-}, BOD, $NH4^+$, and conductivity	*Nitzschia capitellata, Navicula veneta*
9	Downstream	Eutrophication, low altitude	*Cyclostephanos dubius, Thalassiosira pseudonana*
10	Mid-downstream	Slightly disturbed	*Cyclotella polymorpha, Nitzschia acula*
11	Downstream	Moderately Disturbed, low altitude	*Gyrosigma nodiferum, Gyrosigma attenuatum*

those of the original dataset. The similarity was evaluated through component plans, which caused their relative positions to be changed in mirror images (Fig. 7). The dataset selected also showed high correlation of species richness (R^2=0.993) with the original dataset.

There have been several other ways used to reduce data dimensions. The simplest way is to consider the occurrence frequency of species in samples as it is for example, in the statistical software PC-ORD version 4 (MJM Software Design, 2000): users can exclude species fewer than *N* nonzero numbers of occurrences (McCune et al., 2002). However, in this case important species could be excluded from the newly generated dataset. For instance, some species are site specific with low occurrences. In this case, these species are very important to characterize their sampling sites. Therefore, they should be included in the reduced dataset. Another extreme case is that when species are observed in many samples with similar abundance. They do not make a relevant contribution to the characterization of the ecological conditions of their sampling sites because they make similar contributions in all samples. Therefore, they should not be selected as representative species. If we only consider occurrence frequency to select relevant species, very common species occurring in all sample sites are selected as important species, while species characteristically occurring with high abundance at the limited sites may be not selected as important species although they are representative for certain river types. Our method presented in this study does not suffer from these problems because the SI indicates species contribution in the organization of the SOM map, by taking into account occurrences as well as abundances of each species implemented in the connection intensity of SOM. For instance, species *Achnanthes brevipes*, *Cavinula variostriata*, and *Pinnularia acrospheria* showed respectively 33.5, 38.6, and 34.3 of SI, and were chosen in the reduced dataset, although their occurrence frequencies were low with 6, 7, and 5, respectively. Species *A. brevipes* characterized assemblage type 10, *C. variostriata* was assemblage type 1, and *P. acrospheria* was assemblage type 2 (Fig. 6). In contrast, species *Fragilaria delicatissima*, *Nitzschia graciliformis*, and *Pleurosira laevis* showed respectively 19.2, 19.7, and 8.6 of SI, and were not chosen, although their occurrence frequencies were relatively high with 30, 23, and 14, respectively.

As shown in this study the SOM can be considered as an alternative for dimension reduction in the sense that it learns, in an unsupervised way, a map between a 2D lattice and the data space (Carreira-Perpinan, 2001). Through the learning process, the number of reference vectors in the data space is approximately proportional to the data probability density and the map in 2D space is topologically continuous. Although the SOM has been successfully applied in diverse fields including ecological studies, it has also some shortcomings: no cost function to optimize can be defined, no general proofs of convergence exist, and no probability distribution function is obtained. These shortcomings are overcome through trial and error.

The SI indicates the relative importance of species in determining classifications in the SOM. The calculation procedure of the SI is based on the weight values of the SOM. Due to the characteristics of the mathematical formula of the SI, its values depend upon two properties: distribution patterns of species (limited areas or wide areas) and degree of occurrence probabilities of each species (high occurrence (or abundance) species or rare species). Therefore, the SI shows high values when species are observed in limited samples assigned to the same areas in SOM with high occurrence (or abundance). It is very low when a species is observed in many samples in different clusters or in low densities (or occurrence frequencies). The index is highly dependent on the training resolution of the SOM. Therefore, it is also important to choose an optimum SOM map size. The SOM should also be smoothly trained in topology and the map should be optimised. In fact, this approach requires a lot of computation, however computation time is not a critical weakness considering the current speed of processors.

In summary, we propose a method to select relevant species in a large dataset through a self organizing map and structuring index. Through this approach, we built a new dataset with a reduced number of species without losing much information embedded in the original dataset. This computational technique could be applied for preprocessing

data in multivariate analyses and could be useful in ecosystem management needing to reduce the number of variables in large datasets. Our work on dimension reduction could be also helpful for data management and data mining in various other fields of research.

Acknowledgements

This work was supported by the EU projects Rebecca (contract number SSPI-CT-2003-502158) and the Euro-limpacs (contract number GOEC-CT-2003-505540).

REFERENCES

AFNOR, 2000. NFT 90-354: Détermination de l'indice biologique diatomées (IBD). Agence de l'Eau Artois-Picardie, Douai.

Bellman, R., 1961. Adaptive Control Processes: A Guided Tour. Princeton University Press, New Jersey.

Brosse, S., Guegan, J.F., Tourenq, J.N., Lek, S., 1999. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecol. Model. 120, 299–311.

Brown, J.H., 1981. Two decades of homage to Santa Rosalia: toward a general theory of diversity. Am. Zool. 21, 877–888.

Cao, Y., Larsen, D.P., Thorne, RSt-J., 2001. Rare species in multivariate analysis for bioassessment: some considerations. J. N. Am. Benthol. Soc. 20, 144–153.

Carreira-Perpinan, M.A., 2001. Continuous latent variable models for dimensionality reduction and sequential data reconstruction. PhD thesis, University of Sheffield, Sheffield, UK.

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

Céréghino, R., Park, Y.-S., Compin, A., Lek, S., 2003. Predicting the species richness of aquatic insects in streams using a restricted number of environmental variables. J. N. Am. Benthol. Soc. 22, 442–456.

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

European Parliament, 2000. Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for community action in the field of water policy. Off. J. L 327, 1–72.

Fodor, I.K., Kamath, C., 2002. Dimension reduction techniques and the classification of bent double galaxies. Comput. Stat. Data Anal. 41, 91–122.

James, F.C., McCulloch, C.E., 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box? Ann. Rev. Ecolog. Syst. 21, 129–166.

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

Kohonen, T., 2001. Self-organizing maps, Third edition. Springer, Berlin.

Krammer, K, Lange-Bertalot, H. (1986–1991) Bacillariophyceae 1. Teil:Naviculaceae. 876 p.; 2 Teil: Bacillariaceae, Epithemiaceae, Surirellaceae, 596 p.; 3 Teil: Centrales, Fragilariaceae, Eunotiaceae, 576 p.; 4 Teil: Achnanthaceae. Kritische Ergänzungen zu Navicula (Lineolatae) und Gomphonema. 437 p. *In* Süßwasserflora von Mitteleuropa. Band 2/1-4- H. Ettl, J. Gerloff, H. Heynig and D. Mollenhauer (Eds.), G. Fischer verlag., Stuttgart.

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.

Lek, S., Guégan, J.F., 2000. Artificial Neuronal Networks: Application to Ecology and Evolution. Springer-Verlag, Heidelberg.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

Mante, C., Dauvin, J.C., Durbec, J.P., 1995. Statistical-method for selecting representative species in multivariate-analysis of long-term changes of marine communities — application to a macrobenthic community from the bay of Morlaix. Mar. Ecol., Prog. Ser. 120, 243–250.

Mante, C., Durbec, J.P., Dauvin, J.C., 1997. Analysis of temporal changes in macrobenthic communities on the basis of probable species presence. Oceanol. Acta 20, 71–79.

McCune, B., Grace, J.B., Urban, D.L., 2002. Analysis of Ecological Communities. MjM Software Design, Gleneden Beach, OR.

MJM Software Design, 2000. PC-ORD for Windows: multivariate analysis of ecological data, version 4. Gleneden Beach, OR.

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

Palmer, M., 2005. Ordination Methods for Ecologists. [online] http://ordination.okstate.edu/.

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

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

Park, Y.-S., Gevrey, M., Lek, S., Giraudel, J.L., 2005. Evaluation of relevant species in communities: development of structuring indices for the classification of communities using a self-organizing map. In: Lek, S., Scardi, M., Verdonschot, P., Descy, J. P., Park, Y.-S (Eds.), Modelling Community Structure in Freshwater Ecosystems. Springer, Berlin, pp. 369–380.

Preston, F.W., 1962. The canonical distribution of commonness and rarity: part I. Ecology 43, 185–215.

Recknagel, F. (Ed.), 2003. Ecological Informatics: Understanding Ecology by Biologically Inspired Computation. Springer, Berlin.

Snyder, E.B., Robinson, C.T., Minshall, G.W., Rushforth, S.R., 2002. Regional patterns in periphyton accrual and diatom assemblages structure in a heterogeneous nutrient landscape. Can. J. Fish. Aquat. Sci. 59, 564–577.

Stevenson, R.J., 1997. Scale-dependent determinants and consequences of benthic algal heterogeneity. J. N. Am. Benthol. Soc. 16, 248–262.

Tison, J., Giraudel, J.L., Coste, M., Park, Y.S., Delmas, F., 2004. Use of unsupervised neural networks for eco-regional zonation of hydrosystems through diatom communities: case study of Adour–Garonne watershed (France). Arch. Hydrobiol. 159, 409–422.

Tison, J., Park, Y.-S., Coste, M., Wasson, J.G., Ector, L., Rimet, F., Delmas, F., 2005. Typology of diatom communities and the influence of hydro-ecoregions: A study on the French hydrosystem scale. Wat. Res. 39, 3177–3188.

Vesanto, J., 2000. Neural network tool for data mining: SOM Toolbox. Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000). Oulun yliopistopaino, Oulu, Finland, pp. 184–196.

Walley, W.J., Martin, R.W., O'Connor, M.A., 2000. Self-organising maps for classification of river quality from biological and environmental data. In: Denzer, R., Swayne, D.A., Purvis, M., Schimak, G. (Eds.), Environmental Software Systems: Environmental Information and Decision Support. IFIP Conference Series. Kluwer Academic Publishers, Boston Hardbound, pp. 27–41.