Knowledge and Management of Aquatic Ecosystems (2012) 406, 04 http://www.kmae-journal.org

DOI: 10.1051/kmae/2012021

Analysing a large dataset on long-term monitoring of water quality and plankton with the SOM clustering

A. Voutilainen[(1)]*, M. Rahkola-Sorsa[(1)], J. Parviainen[(2)], M.J. Huttunen[(3)], M. Viljanen[(1)]

Received February 20, 2012
Revised June 21, 2012
Accepted August 10, 2012

ABSTRACT

Key-words: *boreal lake, chlorophyll, phytoplankton, Self-Organizing map, water temperature, zooplankton*

The Self-Organizing Map (SOM) proved to be the method of choice for analysing a large heterogeneous ecological dataset. In addition to distributing the data into clusters, the SOM enabled hunting for correlations between the data components. This revealed logical and plausible relationships between and within the environment and groups of organisms. The main conclusions derived from the results were: (i) the structure of early summer plankton community significantly differed from that of late summer community in Lake Pyhäselkä and (ii) plankton community in late summer was characterized by two functional groups. The first group was formed mainly by phytoplankton, rotifers, and small cladocerans, such as *Bosmina* spp., and driven by water temperature. The second group was formed by small copepods and the abundant generalist herbivorous cladocerans *Daphnia cristata* and *Limnosida frontosa*, which, in turn, associated with chlorophyll *a* concentration. Biomasses of *Bosmina* spp. and *D. cristata* showed decreasing monotonic trends during a 20-year study period supposedly due to oligotrophication. Versatile possibilities to cluster data and hunt for correlations between data components offered by the SOM decisively helped to reveal associations across the original variables and draw conclusions. The results would have been undetectable solely on the basis of unorganised values.

RÉSUMÉ

Analyse d'un large ensemble de données de surveillance à long terme de la qualité de l'eau et du plancton par classification SOM

Mots-clés : *lac boréal, phytoplancton, carte auto-organisatrice,*

La carte d'auto-organisation (SOM) s'est avérée être la méthode de choix pour l'analyse d'un large ensemble de données hétérogènes écologiques. En plus de distribuer les données en grappes, la SOM permet la recherche de corrélations entre les composantes des données. Cette étude a révélé des relations logiques et plausibles entre et au sein de l'environnement et des groupes d'organismes. Les principales conclusions tirées des résultats étaient les suivantes : (i) la structure de la communauté planctonique du début de l'été diffère significativement de celle de la communauté de fin de l'été dans le lac Pyhäselkä et (ii) la communauté planctonique en fin d'été a été marquée par deux groupes fonctionnels. Le premier

(1) Department of Biology, University of Eastern Finland, Joensuu campus, PO Box 111, 80101 Joensuu, Finland
(2) Department of Environmental Science, University of Eastern Finland, Kuopio campus, PO Box 1627, 70211 Kuopio, Finland
(3) School of Forest Sciences, University of Eastern Finland, Joensuu campus, PO Box 111, 80101 Joensuu, Finland
* Corresponding author: ari.voutilainen@uef.fi

Article published by EDP Sciences

température de l'eau, zooplancton

groupe a été formé principalement par le phytoplancton, les rotifères, et les petits cladocères, tels que *Bosmina* spp., conditionné par la température de l'eau. Le deuxième groupe a été formé par de petits copépodes et l'abondance des herbivores généralistes cladocères *Daphnia cristata* et *Limnosida frontosa*, associé à la concentration de chlorophylle. Les biomasses de *Bosmina* spp. et *D. cristata* ont montré au cours d'une période d'étude de 20 ans des tendances monotones décroissantes supposées dues à l'oligotrophisation. Les multiples possibilités de classement des données et la recherche de corrélations entre les composants des données offertes par la SOM a bien permis de révéler les associations entre les variables et de tirer des conclusions. Les résultats auraient été indétectables uniquement sur la base de valeurs non organisées.

INTRODUCTION

Analysing of large ecological datasets is a statistical challenge. Researchers often have to deal with problems arising from asymmetric distribution of variables (skewness), autocorrelation, non-random sampling, varying group sizes and missing values, for instance. Here we present a dataset on long-term monitoring of water quality, phytoplankton and zooplankton of a large Finnish boreal lake, Lake Pyhäselkä. The monitoring was originally due to one of the largest pulp mills in the world, Enocell, and the city of Joensuu, which wanted to control their own nutrient loadings into the lake and compare them to diffuse nonpoint source loading.

The Self-Organizing Map (SOM) (Kohonen, 2001) was used to reveal associations between the environment and organisms as well as between and within groups of organisms representing different trophic levels. The SOM is an unsupervised artificial neural network especially feasible in an exploratory data analysis. The SOM can be used as a data visualization tool that performs a non-linear projection from a high-dimensional feature space onto a two-dimensional map lattice (Pölzlbauer *et al*., 2006). In practice, large multivariable data can be presented to the user as intuitive pictures *i.e.* feature maps, which are rather easy to analyse.

A common procedure in aquatic ecology is to cluster the data according to environmental variables and then fit biological variables into the clusters. In the present study, we explored the data by the SOM and let the "dependent variables" (phytoplankton and zooplankton) themselves to search for factors that best explained variability in their abundances. In other words, a so-called training mask for the SOM was constructed from these "dependent variables". The reverse procedure used helps in understanding specifically which associations between environmental factors and biological variables as well as within biological variables are the most affective in a particular context (see also Park *et al*., 2003).

We wanted to use expressly the SOM clustering because the data in question have proven to be too complex for many conventional methods. Previous to the present study we have managed to analyse the data in scattered parts (e.g., Voutilainen and Huuskonen, 2010) but never as a coherent set. The SOM allowed as to project more complicated hypotheses and to build a matrix which included entire data. The study had two main goals: 1) to evaluate suitability of SOM clustering for analysing a large, strongly heterogeneous limnological dataset and 2) to find ecologically relevant cross-trophic level relationships between the environment and organisms. We hypothesised that changes in environmental factors, such as increasing water temperature and decreasing nutrient concentrations i.e. oligotrophication, would affect crustacean zooplankton community mainly indirectly via their direct effects on phytoplankton.

MATERIALS AND METHODS

> *STUDY AREA*

Lake Pyhäselkä is situated in the boreal region in eastern Finland (Figure 1). The lake is humic (water colour 84 mg $Pt{\cdot}L^{-1}$ in 2010) and large with a surface area of 263 km^2, but quite

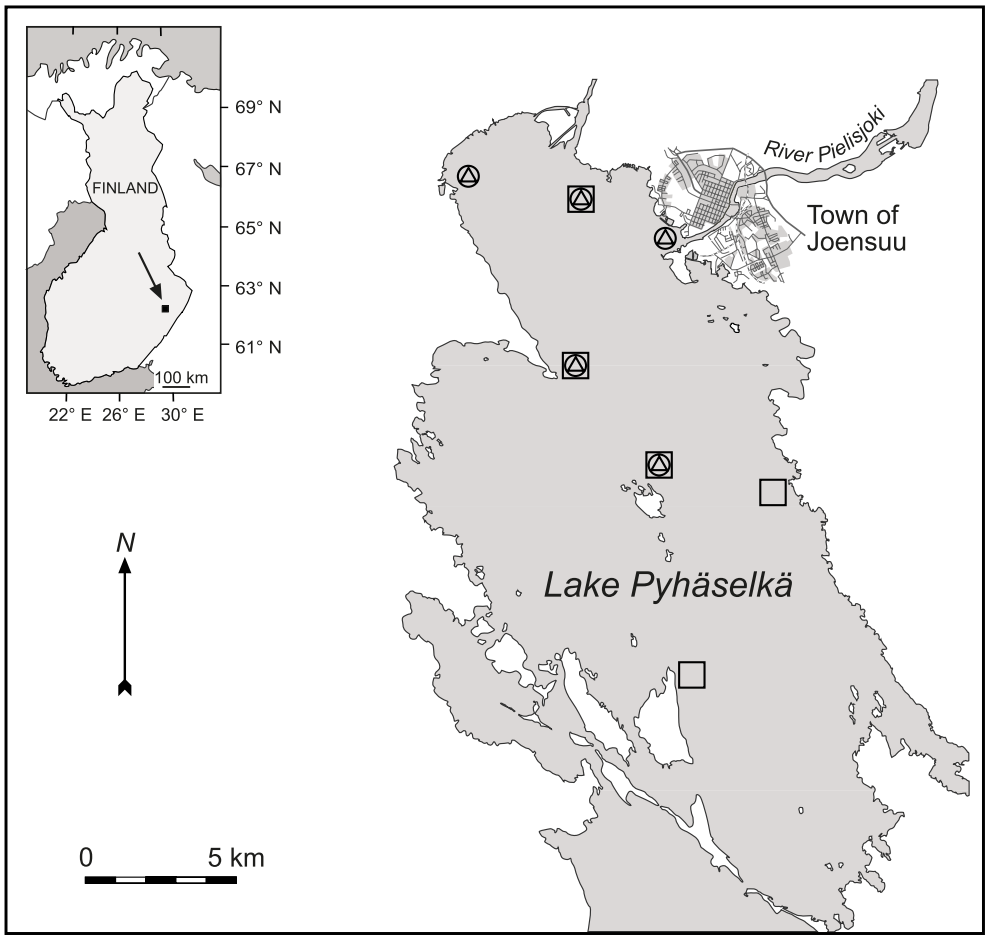

Figure 1
Location of Lake Pyhäselkä in Finland and those of water (triangle), phytoplankton (circle) and zooplankton (square) sampling sites in the lake.

shallow with a mean depth of 9 m. The originally oligotrophic lake became eu-mesotrophic in the 1970s due to nutrient inputs from the Enocell pulp mill and the city of Joensuu. The city lies in the vicinity of the lake. The period of moderately high nutrient loading into Lake Pyhäselkä lasted nearly two decades. In years when the nutrient emission had its maximum (1977–1978), the average total phosphorus (TP) and total nitrogen (TN) concentrations in the lake water were 21 and 586 $\mu g \cdot L^{-1}$, respectively. The plants that treat pulp mill effluents from Enocell and municipal sewage from the city of Joensuu were modernised in 1987 and 1992, respectively, which decreased considerably the nutrient loadings into Lake Pyhäselkä in the 1990s and helped the lake to start recovering towards its pristine water quality (Voutilainen and Huuskonen, 2010). Nowadays, the lake is oligotrophic again (TP 11 $\mu g \cdot L^{-1}$ and TN 444 $\mu g \cdot L^{-1}$ in 2010). Phosphorus is the major nutrient limiting primary production in the lake (Pietiläinen and Niinioja, 2000). No oxygen depletion exists in the lake.

Boreal lake biota typically includes species that are highly specialised to live in well-oxygenated cold water, such as many salmonid fishes, the planktonic crustacean *Mysis relicta*, and cold stenothermic calanoid copepods, such as *Limnocalanus macrurus*, whereas some others are much more ubiquitous having broader thermal tolerances. In any case, however, the species have to tolerate low temperatures as lakes in the boreal region are covered with ice every winter and water temperature under the ice ranges between 0 and 4 °C. In summertime, temperature in the lake uppermost water layers rises up to 20 °C or, occasionally, even higher up to 25 °C. Due to high difference in water temperature between winter and summer life in boreal lakes is cyclic.

> SAMPLING PROCEDURES

Samples for water quality analyses have been taken from Lake Pyhäselkä from five permanent sampling sites (Figure 1) since 1961. For a more detailed description of water sampling and analyses see Pietiläinen and Niinioja (2000) and Voutilainen and Huuskonen (2010). Analyses of phytoplankton primary production (standard SFS 3049) and chlorophyll *a* concentration (SFS 5772) have been carried out continuously since 1973 and 1977, respectively. Samples for phytoplankton have been taken since 1987. The phytoplankton sampling ceased for 10 years (1993–2002), but then began again. Samples for zooplankton have been taken since 1989 to the present without any breaks.

Phytoplankton samples were collected from the uppermost water layer (0–2 m) from five sampling sites (Figure 1) using a Ruttner-type water sampler. Biomasses (fresh weight $\mu g \cdot L^{-1}$) of Cyanophyta (blue-green algae), Cryptophyta (cryptomonads), Chrysophyceae (golden algae), Diatomophyceae (diatoms) and Chlorophyta (green algae) were microscopically estimated from the samples preserved in 0.5% acid Lugol's solution (Holopainen *et al*., 2008). Zooplankton composite samples were collected from the littoral (water depth 0–2 m) and pelagic zones from five sites (Figure 1) approximately once per month from May to October using the Limnos tube sampler with a height of 1 m and volume of 6.7 L (Karjalainen *et al*., 1996a). The littoral sampling sites were mostly rocky shores having only minor vegetation. Therefore, the sites were vulnerable to wind and water currents. Zooplankton samples were taken not only from the pelagic zone, which is a traditional procedure, because it was assumed that the littoral zooplankton would be a better marker of diffuse nonpoint source loading. The zooplankton samples were first concentrated by pouring them through a 50 μm net and preserved with ethanol in the field. Preservation of the samples was assured by adding formaldehyde into sample bottles later in the laboratory. The samples were then pooled over the sampling sites and finally zooplankters were identified microscopically to the lowest taxonomic level possible. In the present study, the zooplankton abundance will be given in biomass ($\mu g\, C \cdot L^{-1}$, see Rahkola *et al.*, 1998). Ranges of physicochemical and biological variables in Lake Pyhäselkä within the studied time periods are shown in Table I and Figure 2.

> STATISTICAL ANALYSES AND MODELLING

The SOM can be understood as a regular low-dimensional (1D or 2D) grid that consists of nodes called neurons (Vesanto, 1999). The size of the neuron grid is roughly determined by the size of the data – a larger dataset requires a larger SOM network. The SOM process is referred to training of the network using "competitive learning". This means that the neuron vectors compete in how to represent the original data most accurately. On each training step, a data sample corresponding to a row in the original data matrix is selected and the best-matching unit (BMU) i.e. neuron is search for it. During the training, the SOM behaves like a flexible net, which folds in a cloud formed by the input data (Vesanto, 1999). As a result, the neuron vectors of the SOM represent groups of the original measurement vectors. Further analyses are now easier because the amount of data is reduced and patterns in the data are more evident. In an exploratory data-analysis, the interpretation of SOM is regularly based on illustrations, where quantitative information is depicted as a colour values on the map lattice (Pölzlbauer *et al*., 2006). Those maps, called component planes, reveal the distribution of each variable of original data on the SOM. The way to look into the SOM is to visualize the results as Unified Distance Matrix (aka *U*-matrix), which illustrates the distances between prototype vectors and provides a general view about the clustering structure (Ultsch and Siemon, 1990). There are several in-depth descriptions about the SOM algorithm (e.g., Kohonen, 2001; Haykin, 2009) and numerous articles on ecological applications of the SOM (e.g., Park *et al*., 2003; Compin and Céréghino, 2007).

The data matrix introduced into the SOM was constituted of 163 rows and 32 columns. Each row represented one week and each column an environmental or biological variable. The variables have been listed in Table I. Since in most cases samples taken within the same

Table I
Variables measured in Lake Pyhäselkä in 1987 – 2009; n indicates how many times samples were taken for the variable in question.

Variable	*n*	Mean	SD	Range
Water temperature (°C)	163	14.1	4.3	3.5–21.6
Precipitation (mm·wk^{-1})	163	23.2	30.5	0–179
Total phosphorus concentration (μg·L^{-1})	162	14	3	8–23
Total nitrogen concentration (μg·L^{-1})	157	446	52	232–643
Chlorophyll *a* concentration (μg·L^{-1})	134	5.2	1.9	0.8–14
Water colour (mg Pt·L^{-1})	131	73	13	40–120
pH	131	6.6	0.2	6.1–7.2
Electrical conductivity (mS·m $^{-1}$)	102	3.7	0.6	2.7–5.4
Phytoplankton production (mg C·m^{-3}·d^{-1})	87	94	67	3–352
Phytoplankton biomass (μg·L^{-1})	59	533	271	185–1362
Chlorophyta	64	26	49	1–375
Chrysophyceae	64	40	44	5–210
Cryptophyta	64	105	91	11–616
Diatomophyceae	64	235	187	27–924
Cyanophyta	60	39	40	0–144
Rotatoria biomass (μg C·L^{-1})	119	15.4	12.8	0.38–75.3
***Asplancha* spp.**	119	1.10	2.02	0–12.0
Cladocera biomass (μg C·L^{-1})	119	50.9	53.0	0.61–305
***Bosmina* spp.**	119	23.0	32.5	0.42–200
***Chydorus* spp.**	119	0.17	0.48	0–3.79
Holopedium gibberum	119	0.89	1.94	0–12.4
Polyphemus pediculus	119	2.24	9.31	0–80.1
Daphnia cristata	119	18.9	22.7	0.03–135
Limnosida frontosa	119	2.67	3.97	0–21.6
Leptodora kindti	119	0.64	1.03	0–6.62
Copepoda biomass (μg C·L^{-1})	119	19.4	10.9	0.41–52.7
***Eudiaptomus* spp.**	119	4.80	4.11	0.08–21.7
***Cyclops* spp.**	119	1.19	1.52	0–6.12
***Meso-* & *Thermocyclops* spp.**	119	7.72	5.44	0.22–25.8
Heterocope appendiculata	119	3.14	4.20	0–19.2
Limnocalanus macrurus	119	0.35	0.69	0–3.34
Eurytemora lacustris	119	0.69	0.83	0–4.98

week were pooled over the sampling sites before analysing them, we were forced not to use a sampling site but a sampling date as "a primary data unit". Moreover, we wanted to see if the SOM would be able to cluster the data according to the plankton annual succession cycle taking place in Lake Pyhäselkä. Variables measured within the same week in the same year were combined so that they represented the same data unit in the analyses. A training mask for the SOM was constructed from biological variables describing the biomasses of total phytoplankton and main zooplankton groups (Rotatoria, Cladocera and Copepoda). The species data were $\log_{10}$ transformed to correct for potential bias caused by skewness. The entire data were normalized to unit variance prior to the modelling. Otherwise the variables with large absolute values would have dominated the map forming.

The hexagonal lattice was selected as the SOM topology type and the batch training procedure was used in the map training. A training mask was utilized to exclude environmental variables from the SOM training so that search for the BMU was based only on the main biological variables (biomasses of phytoplankton and zooplankton). The correct map size (number of neurons) was determined by minimizing the map quantization and topographic errors. The component planes were illustrated with the mean actual values in each map unit. If the map unit was empty for the variable in question, the average value of adjacent map units was used to fill the particular unit.

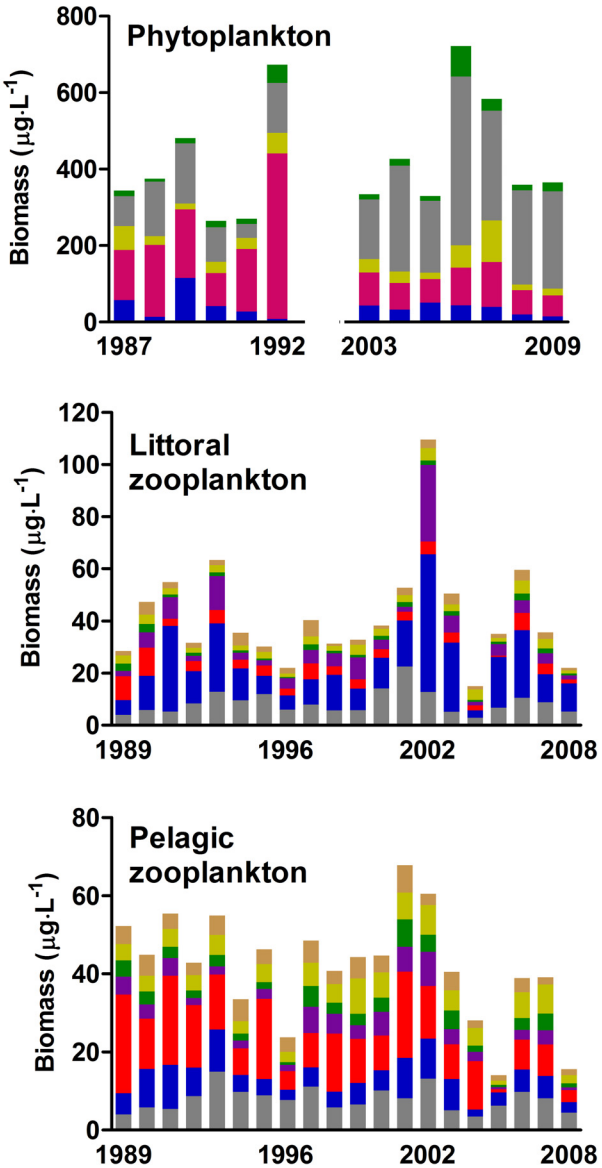

Figure 2
Average annual biomasses of main phytoplankton and zooplankton groups/species in Lake Pyhäselkä, the groups have been indicated as follows. Phytoplankton: Cyanophyta (blue), Cryptophyta (red), Chrysophyceae (yellow), Diatomophyceae (gray), Chlorophyta (green). Zooplankton: rotifers (gray), Bosmina spp. (blue), Daphnia cristata (red), other cladocerans (violet), Eudiaptomus spp. (green), Meso- and Thermocyclops spp. (yellow), other copepods (brown).

The biological variables were chosen mainly on the basis of our previous studies. *Bosmina* spp., *Holopedium gibberum* and *Daphnia cristata* are small- to medium-sized herbivorous cladocerans. In Lake Pyhäselkä, they are numerous and can be considered as the dominant cladoceran species (Karjalainen *et al.*, 1996b; Viljanen *et al.*, 2009). *Eudiaptomus* spp. in turn are small calanoids grazing on phytoplankton and playing a significant role in the lake food web (Karjalainen *et al.*, 1996b; Viljanen *et al.*, 2009). Adults of *Mesocyclops leuckarti* and *Thermocyclops oithonoides* together with *Polyphemus pediculus* are the most common small zooplankton predators (Karjalainen *et al.*, 1996b; Viljanen *et al.*, 2009). *Asplancha* spp. are large rotifers consuming mainly other rotifers, but occasionally also small cladocerans, such as *Bosmina* spp. (Matveeva, 1989).

The trained SOM map was distributed into clusters on the basis of *U*-matrix algorithm and *k*-means method. In the case of *k*-means method, the maximum of Simple Structure Index

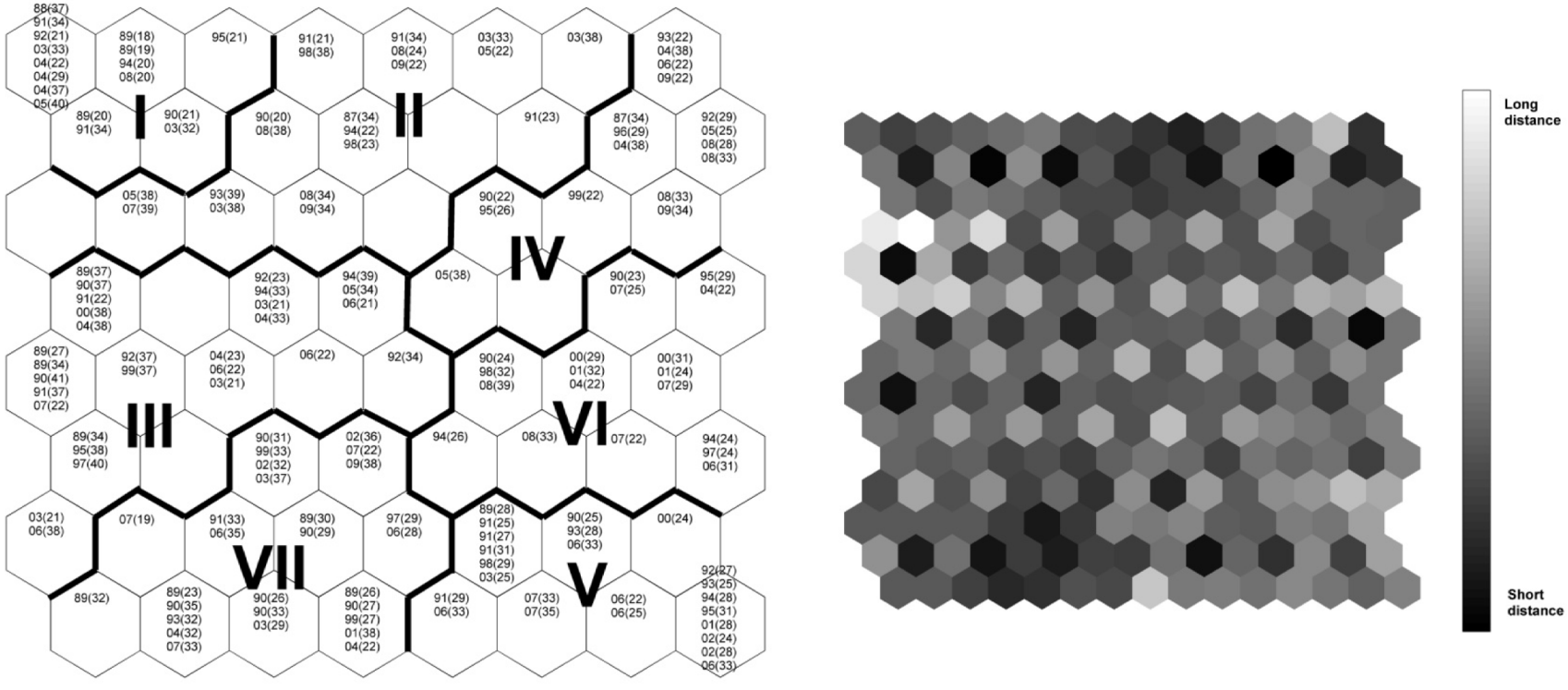

Figure 3

Clustering of the trained SOM units (on the left). The U-matrix and k-means methods were applied to set boundaries on the SOM map. The Latin numbers (I-VII) display clusters, and the codes in each unit of the map represent the sampling date so that 94(39) refers to year 1994 and week 39, for instance. In the U-matrix (on the right), the map units having short distances to the neighbouring units have been signified with dark and thus possible clusters can be seen as dark areas surrounded by light borders.

(SSI) was used to indicate the best partitioning solution in terms of number of clusters (Borcard *et al.*, 2011). In the *U*-matrix (Figure 3), the map units having short distances to the neighbouring units have been signified with dark and thus possible clusters can be seen as dark areas surrounded by light borders. Moreover, the SOM component planes were used for "correlation hunting" (Vesanto, 1999). To do this, a covariance matrix between all component planes was calculated. Each row of this matrix corresponded to one component plane. The matrix was used to train a SOM and each component plane was then placed on the map corresponding to its BMU. If more than one plane had the same BMU, the worst matching of them was moved to its next-best-matching unit. This was repeated until only one component plane located in one unit. The SOM modelling was conducted in Matlab (The Mathworks, Natick, USA) with the SOM Toolbox 2.0 developed by the Laboratory of Information and Computer Science at the Helsinki University of Technology (http://www.cis.hut.fi/projects/somtool-box/).

The non-parametric Mann-Kendall trend test was used to test for monotonic trends in time series of particular variables, which appeared to be especially interesting on the basis of the SOM analyses. The trend analyses were computed using functions for the R statistical language (R 2.11.1, http://www.r-project.org/).

RESULTS

Size of the optimal SOM proved to be 64 neurons corresponding to an 8×8 matrix having a final quantization error of 0.339 and a final topographic error of 0.129. The trained SOM map was distributed into seven clusters on the basis of *U*-matrix algorithm and *k*-means method (Figure 3). The 7-cluster partitioning resulted in the SSI of 0.612.

The cluster I was named as "early summer" as weeks when water temperature and nearly all biological variables had low values clustered in it (Figure 3). The cluster V had characters opposite to the cluster I; water temperature, precipitation, phytoplankton, and *Bosmina* spp., *H. gibberum*, *P. pediculus*, and *Leptodora kindti* of cladocerans showed high values in this cluster. The cluster V corresponded rather well with the results of correlation hunting when the component planes of Phytoplankton, Rotatoria, Cladocera, Diatomophyceae, and *Bosmina* spp. grouped together ("green group" in Figure 4). The cluster VII was termed as "the common ones" as *Eudiaptomus* spp., *M. leuckarti*, *T. oithonoides*, *Eurytemora lacustris*, *D. cristata*, *Limnosida frontosa*, and *Chydorus* spp. had high values in it. In terms of abundance, these species can be considered as the dominants in Lake Pyhäselkä, together with *Bosmina* spp. The cluster VII corresponded well with the group formed by the component planes of Copepoda, *Eudiaptomus* spp., small cyclopoids including *M. leuckarti* and *T. oithonoides*, *L. frontosa*, and *D. cristata* in correlation hunting ("blue group" in Figure 4). Chlorophyll *a* concentration also showed high values in the cluster VII.

The clusters II, III, IV, and VI were not as plausible as the before-mentioned clusters I, V, and VII. The cluster IV was characterized by low precipitation, electrical conductivity, phytoplankton primary production, biomasses of Cryptophyta and Chrysophyceae, and high values of water colour and biomass of *Cyclops* spp. In correlation hunting (Figure 4), precipitation grouped with conductivity ("violet group"), while water colour grouped with *Cyclops* spp., *Limnocalanus macrurus*, and total nitrogen concentration ("red group"). In general, values of water quality and plankton variables increased from the upper left corner to the lower right corner in the trained SOM map.

The following results were derived from the unorganised data. Total fresh biomass of phytoplankton in Lake Pyhäselkä (1987–1992, 2003–2009) varied between 265 and 721 $\mu g \cdot L^{-1}$ (425 ± 148 $\mu g \cdot L^{-1}$, mean ± SD) (Figure 2). The most abundant phytoplankton group was diatoms with a mean biomass of 191 $\mu g \cdot L^{-1}$. The average abundance of diatoms was nearly 2.5 times higher (264 $\mu g \cdot L^{-1}$) in the second monitoring period (2003–2009) compared to that (106 $\mu g \cdot L^{-1}$) in the first period (1987–1992). Total biomass of zooplankton in the littoral zone of Lake Pyhäselkä (1989–2008) varied between 15.1 and 110 $\mu g \cdot L^{-1}$ (41.8 ± 20.5 $\mu g \cdot L^{-1}$, mean ± SD) (Figure 2). Cladocerans (26.3 ± 17.7 $\mu g \cdot L^{-1}$, mean ± SD), rotifers (8.60 ± 4.60 $\mu g \cdot L^{-1}$,

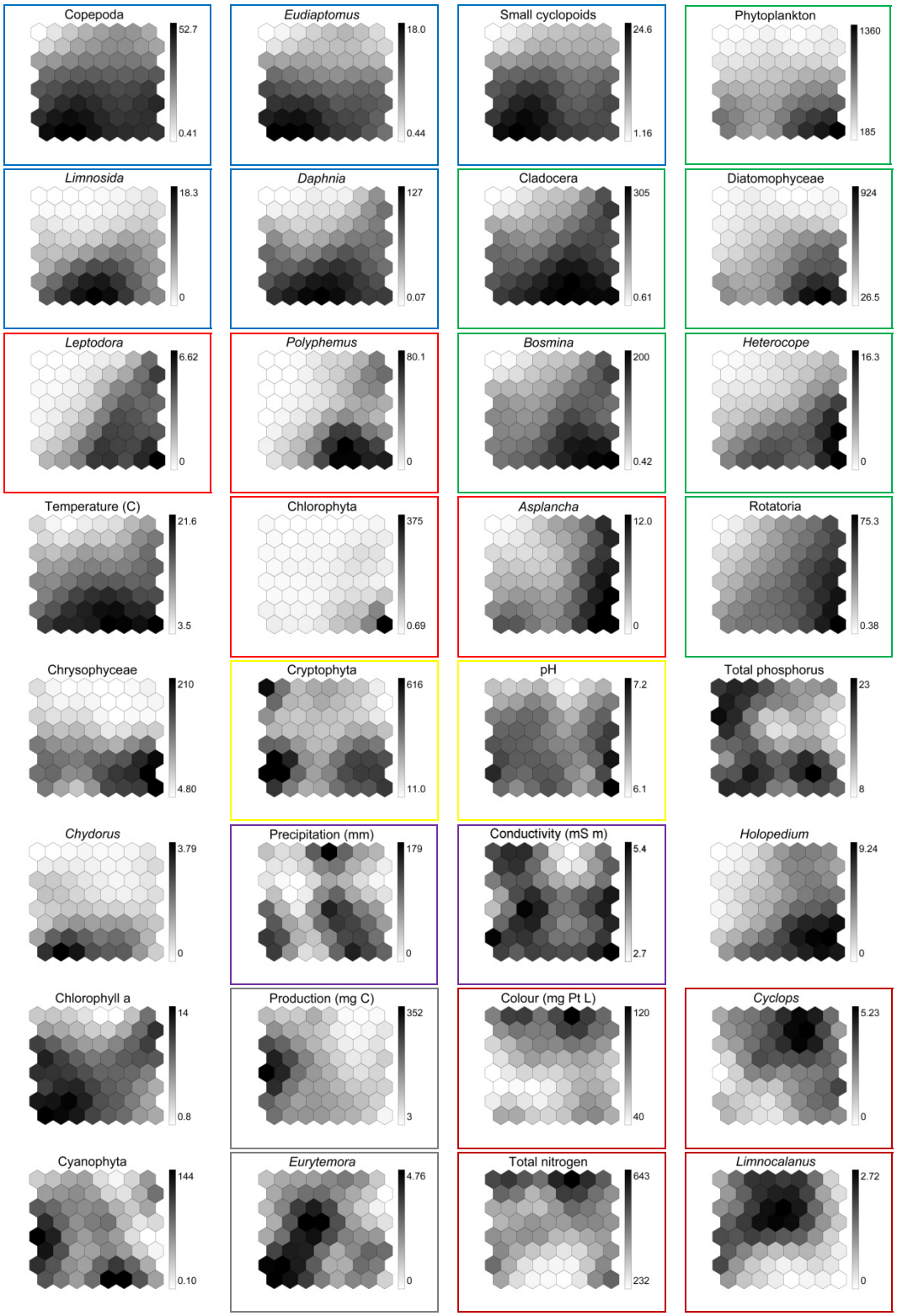

Figure 4

Visualization of environmental and biological variables on the trained SOM map. The variables were grouped by correlation hunting, and the variables belonging to the same group are surrounded with a line of a certain colour. Not all variables belonged to the groups. In the planes, dark represents high values, while light is low. The unit is $\mu g \cdot L^{-1}$ *unless otherwise mentioned.*

mean ± SD) and copepods (6.94 ± 2.63 μg·L^{-1}, mean ± SD) constituted 63, 21 and 17% of the total zooplankton biomass, respectively. Total abundance of zooplankton in the pelagic zone of Lake Pyhäselkä (1989–2008) varied between 14.0 and 67.8 μg·L^{-1} (42.1 ± 14.1 μg·L^{-1}, mean ± SD) (Figure 2). Like in the littoral zone, cladocerans (22.1 ± 10.1 μg·L^{-1}, mean ± SD) were the most abundant zooplankton group. They constituted 53% of the total zooplankton biomass. Rotifers (7.88 ± 3.02 μg·L^{-1}, mean ± SD) and copepods (12.0 ± 4.33 μg·L^{-1}, mean ± SD) constituted 19 and 29% of the total biomass, respectively.

A detailed analysis of temporal changes in Lake Pyhäselkä revealed that water temperature (Mann-Kendall, $\tau = 0.146$, $P < 0.05$) and biomass of rotifers ($\tau = 0.134$, $P < 0.05$) showed increasing linear trends in Lake Pyhäselkä in 1989–2008. Biomasses of *D. cristata* and *Bosmina* spp., in turn, showed decreasing linear trends during the same period of time ($\tau = -0.138$, $P < 0.05$ and $\tau = -0.337$, $P < 0.001$, respectively). As the density of *D. cristata* (individuals·L^{-1}) did not show a similar trend ($\tau = -0.049$, $P > 0.05$), the decrease in biomass was not a result of decrease in *D. cristata* density, but due to decrease in average size of *D. cristata*. In the case of *Bosmina* spp., also density showed a strong decreasing trend ($\tau = -0.316$, $P < 0.001$).

DISCUSSION

The present SOM analyses resulted in logical and plausible relationships between the environment and organisms as well as between and within the groups of organisms. Two main ecological conclusions were: (i) the structure of early summer plankton community significantly differed from that of late summer community in Lake Pyhäselkä and (ii) plankton community of Lake Pyhäselkä in late summer was characterized by two functional groups. The conclusions were to some extent in accordance with our earlier results derived by a conventional DCA (Detrended Correspondence Analysis) method (Rahkola-Sorsa, 2008). The species composition of zooplankton assemblage in Lake Pyhäselkä greatly differs between seasons and this seasonal variation overcomes interannual variation (Rahkola-Sorsa, 2008). The first functional plankton group revealed by the SOM analyses was formed mainly by phytoplankton, particularly diatoms, rotifers, small herbivorous cladocerans, such as *Bosmina* spp., and their potential cladoceran predators and driven by water temperature. The second group was formed by abundant smaller copepods together with the efficient filter feeders *D. cristata* and *L. frontosa*, which, in turn, associated with chlorophyll *a* concentration. In general, adequate food supply in terms of high chlorophyll *a* concentration favours both generalist cladocerans, including *Daphnia* spp. and *Limnosida* spp., and copepods, which are more selective feeders. It would have been impossible to draw these conclusions solely on the basis of original values (cf. Karjalainen *et al*., 1996b), without running the SOM and hunting for correlations between the SOM components.

In addition to the main conclusions, we were able to state that Lake Pyhäselkä plankton community has changed so that (i) biomass of rotifers has increased, (ii) that of cladocerans has decreased, and (iii) diatoms have become more abundant during the monitoring period.

Structure of plankton communities is traditionally hypothesised to be determined by environmental factors, such as water temperature and nutrients, along with interactive biotic determinants, such as competition and predation. In most cases, the hypothesis seems to hold true (Shurin *et al.*, 2010), albeit the main determining factor usually depends on season – at least in dimictic boreal lakes. In this study, water temperature came up in the analyses associating positively with biomass of phytoplankton, excluding Cryptophyta, and that of rotifers and certain cladocerans. Both water temperature and rotifer biomass has shown increasing trends in Lake Pyhäselkä since 1989. A combination of trends similar to that detected in Lake Pyhäselkä, increasing water temperature and rotifer biomass and decreasing nutrient concentrations, has been observed in a deep monomictic peri-alpine lake, Lake Geneva, during the period 1969–1998 (Molinero *et al.*, 2006). In this case, the trends were suggested to be a consequence of combining climatic change and anthropogenic actions (Molinero *et al.*, 2006). The increasing trend in Rotatoria biomass of Lake Pyhäselkä was most likely linked to

the increasing trend in water temperature. Although rotifers constitute a highly heterogeneous group of organisms, increasing temperature in general favours them in terms of reproduction and growth (e.g., Galkovskaja, 1987).

Surprisingly, water temperature only weakly associated with copepods in Lake Pyhäselkä; low temperature reduced biomass of small copepods, such as *M. leuckarti* and *T. oithonoides*, but high temperature did not clearly increase it. This is not in accordance with earlier findings concerning especially thermal preference of *M. leuckarti* and *T. oithonoides*, which both are warm-water species (Vijverberg, 1980; Herzig, 1983). On the other hand, biomasses of small copepods related positively to chlorophyll *a* concentration. As nutrient levels in Lake Pyhäselkä have always been quite low indicating oligo- or mesotrophic conditions, despite the high nutrient inputs from municipal and industrial sources especially between the late 1970s and early 1980s (Voutilainen and Huuskonen, 2010), it has been thought that changes in trophic status do not play a leading role in structuring plankton community of Lake Pyhäselkä.

The new SOM results and time series analyses suggested rethinking of plankton dynamics in Lake Pyhäselkä. The observed decrease in the total Cladocera biomass was mainly due to decrease in average size of *D. cristata* and abundance of *Bosmina spp.* The first one is a medium-sized (length range 520–760 μm, own unpublished data) and the second one a small-sized (length range 420–540 μm, own unpublished data) herbivore. Smaller copepods together with the cladocerans *D. cristata* and *L. frontosa* expressly related to chlorophyll *a* concentration, which indicates bottom-up regulation of these consumers. Nowadays, nutrient levels in Lake Pyhäselkä are actually so low that phytoplankton growth most probably is limited by restricted availability of phosphorus (see Pietiläinen and Niinioja, 2000). This can also cause direct reduction in cladoceran growth. Cladocerans, especially *Daphnia* spp., have higher tissue phosphorus content and thus a higher carbon to phosphorus ratio than copepods. This means that cladocerans have a high demand for phosphorus, which makes them potentially vulnerable to phosphorus deficit under oligotrophic conditions (see Sommer and Stibor, 2002). Consequently, the observed decreasing monotonic trends in biomasses of *D. cristata* and *Bosmina* spp. in Lake Pyhäselkä in 1989–2008 may be partly due to phosphorus limitation – as well as due to changes in predation pressure as will be discussed later in this paper. A declining trend in *Bosmina* spp. abundance due to oligotrophication has been observed in a large warm-monomictic lake, Lake Constance, during the period 1979–1998 (Straile and Müller, 2010).

The SOM clustering gave a hint of predator-prey interactions among the zooplankton community, as the predatory cladocerans *L. kindti* and *P. pediculus* associated with their potential prey species, the small cladocerans *Bosmina* spp. (Matveeva, 1989). Biomass of the predatory rotifers *Asplancha* spp. also related to that of *L. kindti* and *P. pediculus*, but it is difficult to state, what does the observed relationship between *Asplancha* spp. and these predatorous cladocerans actually mean. It is possible that the species *Asplancha* spp., *L. kindti*, and *P. pediculus* partially prey on same small zooplankton species, which, in turn, are able to control abundances of their predators. It also is probable that *Asplancha* spp. play a dual role in this game; they prey on *Bosmina* spp. and are preyed on by larger predators (Williamson, 1983; Matveeva, 1989). Relationships between *L. kindti* and its preys are rather complex, as in most cases *L. kindti* seems to be a generalist feeder (Cummins *et al.*, 1969).

Decrease in crustacean zooplankton average size is traditionally related to high predation pressure so that larger-sized species and individuals are eliminated by the predators and then replaced by smaller-sized species and individuals (Brooks and Dodson, 1965). In the case of Lake Pyhäselkä, this would indicate increase in predation by pelagic fish, mainly smelt (*Osmerus eperlanus*), vendace (*Coregonus albula*) and small perch (*Perca fluviatilis*), which all prey on both *Daphnia* spp. and *Bosmina* spp., but much less on other cladocerans, such as *Limnosida* spp. and *Holopedium* spp. (Viljanen, 1983; Sandlund *et al.*, 1987). In Lake Mjoesa, for instance, vendace and smelt preyed on *Daphnia galeata*, not on *D. cristata*, although the latter one was abundant (Sandlund *et al.*, 1987). The abundance of perch, on the other hand, has clearly increased in Lake Pyhäselkä since the 1970s (Voutilainen and Huuskonen, 2010). As juvenile perch prefer large cladocerans ($>$799 μm) as food (Karjalainen *et al.*, 1998),

increase in their abundance will without doubt reflect as strengthened predation pressure on cladocerans. Unfortunately, without further investigations, we are unable to say exactly the preferred prey species of vendace and smelt in Lake Pyhäselkä, and whether abundance of these pelagic zooplanktivores has shown an increasing trend in the lake during the past couple of decades.

The detected increase in biomass of Diatomophyceae is not restricted only to Lake Pyhäselkä, but it is a much more common phenomenon in Finnish lakes (Arvola *et al*., 2011), which, on an average, are oligotrophic (Mitikka and Ekholm, 2003). This is suggested to be a result of markedly decreased discharge of nutrients into Finnish lakes during the last three decades due to improved purification of municipal and industrial wastewaters (Räike *et al*., 2003). Increasing silicon-phosphorus -ratio, in general, favours diatoms (Kalff and Knoechel, 1978).

The SOM is the method of choice for analysing large ecological datasets. In addition to distributing the original data into clusters, correlations between the SOM component planes derived from a trained SOM map resulted in logical associations between and within groups of organisms and the environment. The associations reported would have been impossible to detect solely on the basis of the original unorganised values.

ACKNOWLEDGEMENTS

We wish to thank A.-L. Holopainen, who was responsible for organising the phytoplankton samplings in Lake Pyhäselkä and interpreting the results during the study period 1987–2009. We also thank the diligent and skilful technicians and laboratory workers for taking and analyzing the samples, K. Kyyrönen for drawing the map figure, and two anonymous reviewers for their valuable comments on the manuscript. Financial support provided by the Centre of Expertise in Biology of Environmental stress and Risk Assessment (University of Eastern Finland) and the Academy of Finland (Project 14159) is gratefully acknowledged.

REFERENCES

Arvola L., Järvinen M. and Tulonen T., 2011. Long-term trends and regional differences of phytoplankton in large Finnish lakes. *Hydrobiologia*, *660*, 125–134.

Borcard D., Gillet F. and Legendre P., 2011. Numerical Ecology with R. Springer, New York, 306 p.

Brooks J.L. and Dodson S.I., 1965. Predation, body size, and composition of plankton. *Science*, *150*, 28–35.

Compin A. and Céréghino R., 2007. Spatial patterns of macroinvertebrate functional feeding groups in streams in relation to physical variables and land-cover in Southwestern France. *Landscape Ecol.*, *22*, 1215–1225.

Cummins K.W., Costa R.R., Rowe R.E., Moshiri G.A., Scanlon R.M. and Zajdel R.K., 1969. Ecological energetics of a natural population of the predaceous zooplankter Leptodora kindtii Focke (Cladocera). *Oikos*, *20*, 189–223.

Galkovskaja G.A., 1987. Planktonic rotifers and temperature. *Hydrobiologia*, *147*, 307–317.

Haykin S., 2009. Neural Networks and Learning Machines. Prentice Hall, New Jersey, 906 p.

Herzig A., 1983. The ecological significance of the relationship between temperature and duration of embryonic development in planktonic freshwater copepods. *Hydrobiologia*, *100*, 65–91.

Holopainen A.-L., Lepistö L., Niinioja R. and Rämö A., 2008. Spatiotemporal and long-term variation in phytoplankton communities in the oligotrophic Lake Pyhäjärvi on the Finnish-Russian border. *Hydrobiologia*, *599*, 135–141.

Kalff J. and Knoechel R., 1978. Phytoplankton and their dynamics in oligotrophic and eutrophic lakes. *Ann. Rev. Ecol. Syst.*, *9*, 475–495.

Karjalainen J., Rahkola M., Viljanen M., Andronikova I.N. and Avinskii V.A., 1996a. Comparison of methods used in zooplankton sampling and counting in the joint Russian-Finnish evaluation of the trophic state of Lake Ladoga. *Hydrobiologia*, *322*, 249–253.

Karjalainen J., Holopainen A.-L. and Huttunen P., 1996b. Spatial patterns and relationships between phytoplankton, zooplankton and water quality in the Saimaa lake system, Finland. *Hydrobiologia*, *322*, 267–276.

Karjalainen J., Ollikainen S., Staff S., Viljanen M. and Väisänen P., 1998. Larval fish communities in Lake Puruvesi: species composition and diet (Puruveden kalanpoikasyhteisöt: koostumus ja ravinnonkäyttö). *In*: Grönlund E., Simola H., Viljanen M., Niinioja R. (eds.), Saimaa-seminaari 1998, Saimaa nyt ja tulevaisuudessa, University of Joensuu, Publications of Karelian Institute 122, Joensuu, 52–55.

Kohonen T., 2001. Self-Organizing Maps. Springer-Verlag, Berlin, 501 p.

Matveeva L.K., 1989. Interrelations of rotifers with predatory and herbivorous Cladocera: a review of Russian works. *Hydrobiologia* , 186/187, 69–73.

Mitikka S. and Ekholm P., 2003. Lakes in the Finnish Eurowaternet: status and trends. *Sci. Total Environ.*, *310*, 37–45.

Molinero J.C., Anneville O., Souissi S., Balvay G. and Gerdeaux D., 2006. Anthropogenic and climate forcing on the long-term changes of planktonic rotifers in Lake Geneva, Europe. *J. Plankton Res*., 28, 287–296.

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

Pietiläinen O.-P. and Niinioja R., 2000. Nitrogen and phosphorus as algal growth limiting factors in a boreal lake. *Verh. Internat. Verein. Limnol.*, *27*, 2944–2947.

Pölzlbauer G., Dittenbach M. and Rauber A., 2006. Advanced visualization of Self-Organizing Maps with vector fields. *Neural Networks*, *19*, 911–922.

Rahkola M., Karjalainen J. and Avinsky V.A., 1998. Individual weight estimates of zooplankton based on length-weight regressions in Lake Ladoga and Saimaa Lake system. *Nordic J. Freshw. Res.*, *74*, 110–120.

Rahkola-Sorsa M., 2008. The structure of zooplankton communities in large boreal lakes and assessment of zooplankton methodology. *University of Joensuu, PhD Dissertations in Biology*, 59, 119 p.

Räike A., Pietiläinen O.-P., Rekolainen S., Kauppila P., Pitkänen H., Niemi J., Raateland A. and Vuorenmaa J., 2003. Trends of phosphorus, nitrogen and chlorophyll *a* concentrations in Finnish rivers and lakes in 1975–2000. *Sci. Total Environ.*, *310*, 47–59.

Sandlund O.T., Naesje T.F. and Kjellberg G., 1987. The size selection of *Bosmina longispina* and *Daphnia galeata* by co-occurring cisco (*Coregonus albula*), whitefish (*C. lavaretus*) and smelt (*Osmerus eperlanus*). *Arch. Hydrobiol.*, *110*, 357–363.

Shurin J.B., Winder M., Adrian R., Keller W., Matthews B., Paterson A.M., Paterson M.J., Pinel-Alloul B., Rusak J.A. and Yan N.D., 2010. Environmental stability and lake zooplankton diversity – contrasting effects of chemical and thermal variability. *Ecol. Lett.*, *13*, 453–463.

Sommer U. and Stibor H., 2002. Copepoda–Cladocera–Tunicata: The role of three major mesozooplankton groups in pelagic food webs. *Ecol. Res.*, 17, 161–174.

Straile D. and Müller H., 2010. Response of *Bosmina* to climate variability and reduced nutrient loading in a large lake. *Limnologica*, 2, 92–96.

Ultsch A. and Siemon H.P., 1990. Kohonen's self organizing feature maps for exploratory data analysis. *In*: Proceeding of the INNC'90 International Neural Network Conference, Kluwer, Dordrecht, 305–308.

Vesanto J., 1999. SOM-based data visualization methods. *Intell. Data Anal.*, *3*, 111–126.

Vijverberg J., 1980. Effect of temperature in laboratory studies on development and growth of Cladocera and Copepoda from Tjeukemeer, The Netherlands. *Fresh. Biol.*, *10*, 317–340.

Viljanen M., 1983. Food and food selection of cisco (*Coregonus albula* L.) in a dysoligotrophic lake. *Hydrobiologia*, *101*, 129–138.

Viljanen M., Holopainen A.-L., Rahkola-Sorsa M., Avinsky V., Ruuska M., Leppänen S., Rasmus K. and Voutilainen A., 2009. Temporal and spatial heterogeneity of pelagic plankton in Lake Pyhäselkä, Finland. *Boreal Env. Res.*, *14*, 903–913.

Voutilainen A. and Huuskonen H., 2010. Long-term changes in the water quality and fish community of a large boreal lake affected by rising water temperatures and nutrient-rich sewage discharges – with special emphasis on the European perch. *Knowl. Managt. Aquatic Ecosyst.*, *397*, 03.

Williamson C.E., 1983. Invertebrate predation on planktonic rotifers. *Hydrobiologia*, 104, 385–396.