

Egyetemi doktori (PhD) értekezés tézisei

**NEURÁLIS HÁLÓZATOK ALKALMAZÁSA AZ
ÖKOLOGIAI ADATOK ÉRTÉKELÉSÉBEN**

Várbíró Gábor

Témavezetők: Dr. Tóthmérész Béla

Dr. Borics Gábor



DEBRECENI EGYETEM
Juhász-Nagy Pál Doktori Iskola
Debrecen, 2009

A doktori értekezés előzményei és célkitűzései;

A 2001-ben bevezetett EU Vizkeretirányelv, melynek célja a felszíni és felszín alatti vizek jó ökológiai állapotának elérése 2015-re, több élőlénycsoport rendszeres monitorozását írja elő, mely tevékenység nagy mennyiségű ökológiai adatot eredményez. A rendszeres adatgyűjtés és az így keletkező adatok adatbázisba rendezése nemcsak a strukturált adatfeldolgozást teszi lehetővé, hanem az adatok kiértékelésében is nagy szerepet kap. A benne tárolt adatok élettartama nemcsak a vizsgálat idejére korlátozódik, így azok a későbbi elemzésekben is részt vehetnek.

Az ökológiai folyamatok tér és időbeli összetettsége, a fajok és a környezeti változók közötti nem lineáris kapcsolat és a változók normalitásának hiánya komoly kihívást jelent az adatfeldolgozásban és az ökológiai folyamatok feltárásában. A mesterséges neurális hálózatok vonzó alternatív eszközzé váltak az ökológiai adatelemzés során, mert számos előnyös tulajdonsággal rendelkeznek, mint például a nem-linearitás, adaptivitás, általánosítás és modell függetlenség (nem szükséges a-priori modell). A neurális hálózatok alkalmazása, - bár már az 1950-s években megtörténtek a kezdeti lépések - csak az utóbbi évtizedekben terjedt el és vált az ökológusok által is használt statisztikai eszköztár részévé. A neurális hálózatok erős és hathatós segítséget nyújtottak az ökológiai modellezésben, osztályozási feladatok megoldásában és a mintázatelemzések során (Paisley et al., 2003).

A neurális hálózatokat számos hidrológiai és hidrobiológiai kutatásban sikerrel alkalmazták; így például a vízminőségi paraméterek modellezésében (e.g. Daniell and Wundke, 1993; Lachtermacher and Fuller, 1994; Schizas et al., 1994; Kaluli et al., 1998; Wen and Lee, 1998) vagy az élőlényközösségek és a környezeti változók közötti összefüggések vizsgálatában (e.g. Chon et al., 1996; Lek et al., 1996; Recknagel et al., 1998; Guegan et al., 1998; Lee et al., 1998; Maier et al., 1998).

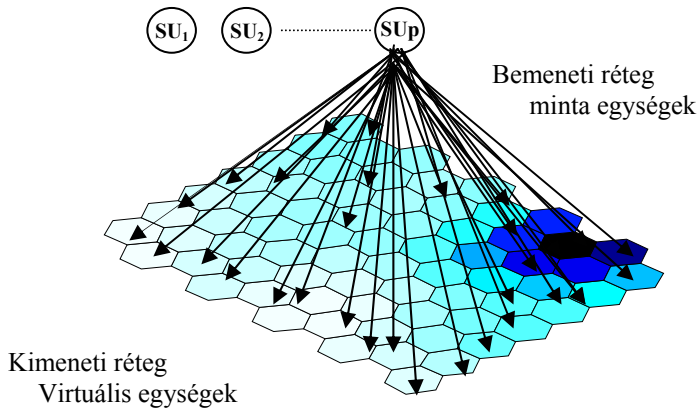
Nagymennyiségű adat elemzésénél a hagyományos többváltozós elemzések számos esetben nehezen értelmezhetők és eredményik nem jeleníthetők meg közérthető formában. A Kohonen-féle

Önszervező Térkép Módszer egy olyan neurális hálózati típus, amely lehetővé teszi nagymennyiségű, összetett, (több dimenziós), adatok elemzését és értékelését. A Teuvo Kohonen által kifejlesztett módszer a dimenzió csökkentése révén a kiindulási adatmátrixot kétdimenziós hexagonális térképre képezi le, megőrizve a kiindulási adatok topológiai és metrikai sajátosságait (Kohonen 2001). A módszer legfőbb erőnye az adatok szemléletes megjelenítése (Vesanto and Alhoniemi 2000).

Anyag és módszer

A statisztikai elemzések elvégzéséhez az önszervező térkép (SOM) algoritmust használtuk. Az algoritmus lehetővé teszi nagy adatbázisok sokdimenziós adatainak leképezését kétdimenziós hexagonális térképpé. Az egymáshoz közeli térkép egységek (neuronok) úgy helyezkednek el a végső térképen, hogy az egymáshoz hasonló tulajdonsággal rendelkező neuronok egymás közelében, míg a különbözőek a térkép távolabbi pontjain helyezkednek el. A SOM a mintákat és a változókat egyidejűleg is képes megjeleníteni, s ugyanakkor a változók fontosságát is meghatározhatjuk. (Kohonen, 2001). Az értékelés során a SOM Toolbox modult (<http://www.cis.hut.fi/projects/somtoolbox>) és a MATLAB™ fejlesztői környezetet használtuk.

A SOM neurális hálózat a nem-ellenőrzött tanulási típusú hálózatok kategóriájába tartozik, ami azt jelenti, hogy nem adunk meg kimeneti paramétert, a csoportok képzését az algoritmus önnállóan végzi. Maga a hálózat egy bemeneti (input) és egy kimeneti (output) rétegből áll. A bemeneti réteg tartalmazza a mintaegységeket (SU), a kimeneti réteg a virtuális egységekből (VU) vagy neuronokból áll. A kimeneti réteg neuronjai a bemeneti réteg súlyvektorainak megfelelő mintázatot vesznek fel egy tanulási folyamat során. Ez ökológiai analógiára lefordítva annyit tesz, hogy a kiindulási mintavételekhez (SU) tartozó faj-abundancia viszonyokra jellemző lesz a végső SOM térkép neuronjainak súlyfaktora, azaz minden egyes neuron egy egy virtuális faj abundancia listával fog rendelkezni. Az eljárás kellően robusztus, lehetővé teszi több ezer mintavételi egység sűrítését néhány virtuális neuronba. A végső SOM térkép leggyakoribb ábrázolási módja a hexagonális rács (1. ábra).



1. ábra. A SOM neurális hálózat felépítése

A SOM algoritmusát az alábbiak szerint adhatjuk meg:

- 1. lépés: Ciklus $t=0$, a virtuális egységet (VU_p) inicializáljuk, a súlyokat (w) az input adatokból töltjük fel.
- 2. lépés: Egy mintaegységet kiválasztunk (SU_j) véletlenszerűen
- 3. lépés: Egy megfelelő metrika segítségével meghatározzuk a SU_j és az összes VU_p közötti távolságot.
- 4. lépés: A SU_j -hez legközelebbi VU_c lesz a győztes neuron, ez a neuront Legjobban hasonlító egységnek (BMU) nevezzük
- 5. lépés: A virtuális egységek VU_p súlyfaktorai a tanulási szabály alapján módosulnak, úgy, hogy az új faktorok a SU_j bemeneti vektorához hasonlítsanak.
- 6. lépés: Újabb ciklus kezdődik $t+1$, amíg $t < t_{\max}$ addig 2. lépésre lép; a ciklus akkor fejeződik be, ha eléri a kívánt számú ismétlést.

Az első lépésben a felhasználó dönti el a térkép méretét. A kimeneti réteg neuronjainak optimális száma az alábbiak szerint számolható (Vesanto, 2000):

$$Nn = 5\sqrt{SU_d} \text{ (eq. 1.)}$$

Ahol Nn a neuronok száma a végső térképen és SU_d a mintavételi egységek száma.

A 3. lépésben a leggyakrabban használt távolság metrika az Euklideszi.

$$d(x, w_i) = \left[\sum_{j=1}^p (x_j - w_{ij})^2 \right]^{\frac{1}{2}}$$

Euklideszi távolság (eq. 2.)

$$w_{ik}(t+1) = w_{ik}(t) + h_{ck}(t)[x_{ij}(t) - w_{ik}(t)]$$

Euklideszi tanulási szabály (eq. 3.)

Az 5. lépésben Euklideszi tanulási szabály $h_{ck}(t)$ változója a szomszédsági függvény. A 4. lépésben kiválasztott BMU neuron nem az egyetlen, amelynek vektorait módosítjuk, hanem a szomszédos neuronok is módosulnak. A leggyakrabban használt szomszédsági függvény a Gauss féle:

$$h_{ck}(t) = \alpha(t) \cdot \exp\left(-\frac{\|r_k - r_c\|^2}{2\sigma^2(t)}\right)$$

(eq. 4.)

$\|r_k - r_c\|$: az Euklideszi távolság a győztes neuron VU_c és a térkép többi neuronja VU_k között,

σ : időben csökkenő faktor azt határozza meg, hogy a térkép mekkora része érintett a tanulási folyamatban,

α : a tanulási faktor. Mindkét faktor idővel csökken, és nullához konvergál.

A SOM jellemzésére és a karakterfaj elemzések során a strukturális indexet (SI) használtuk. Az index azt mutatja meg, hogy az egyes változók (súlyfaktorok) milyen mértékben felelősek a SOM térkép topológiájáért (Park et al., 2005). Azok a változók (esetünkben fajok) amelyek nagyon erős grádiens eloszlást mutatnak, nagy SI index

értékkel rendelkeznek, karakterfajoknak tekinthetők. Az SI értékét az alábbiak szerint számolhatjuk:

$$SI_i = \sum_{j=1}^S \sum_{k=1}^{j-1} \frac{|W_{ij} - W_{ik}|}{\|r_j - r_k\|}$$

(eq. 5.)

ahol w_{ij} és w_{ik} az i^{th} faj súlyfaktorai,

$\|r_j - r_k\|$ a topológiai távolság a térkép két j és k neuronja között,

S a SOM térkép neuronjainak száma.

Az algoritmus eredményeképpen kapott térkép neuronjait a kiindulási mintaegységekkel megfeleltethetjük és ábrázolhatjuk. Bár az egymáshoz hasonló minták egymás közelében helyezkednek el, a mintacsoportok elkülönítése nem minden esetben egyértelmű.

A mintacsoportok elkülönítésére számos lehetőségünk van. Használhatunk hagyományos klaszterezési technikát. Mi az elemzések során a szakirodalom által ajánlott k -átlag módszert használtuk. A megfelelő clusterszám meghatározása a Davies-Bouldin indexet felhasználásával történt (Davies and Bouldin, 1979)

Az ökológiai adatbázis az adatok illetve rekordok stukturált összessége. Napjainkban a leggyakrabban használt adatbázis modell a relációs adatmodell. A reláció nem más, mint egy táblázat, a táblázat soraiban tárolt adatokkal együtt. A relációs adatbázis, pedig a relációk összessége. Ennek alapstruktúrája az adattábla, amely sorokból (rekordok) és oszlopokból (mezők) áll. A relációk vagy kapcsolatok az adatbázis táblái között állnak fenn. A táblák oszlopai vagy mezői az adatok tulajdonságai, pl. mintavételi hely neve, mintavétel dátuma, a sorok vagy adatok pedig a konkrét értékeket tartalmazzák pl. 2009.05.27., Concó, Ács

A relációs adatbázis előnye, hogy minimalizálja a tárolt redundáns információt, ezáltal csökkenti az adatbázis méretét és az egyszerre kezelt adatok mennyiségét, valamint a kapcsolatok révén bármely mező szerinti lekérdezést ill. csoportosítást lehetővé tesz.

Eredmények

A disszertáció öt tudományos közleményt tartalmaz. A disszertációban részletesen bemutattam a SOM algoritmusát,

valamint az ökológiai adatbázis szerkezetét. A megfelelő adatszerkezet és a SOM statisztikai módszer segítségével lehetővé vált magyarországi folyóvíz típusok jellemzése a fitoplakton közösségek alapján.

Létrehoztam egy fitoplakton és egy fitobenton adatbázist amelyet jelenleg is használnak a Környezetvédelmi Felügyelőségek és ezek részei az országos monitoring hálózatnak is.

A SOM módszer segítségével jellemeztem a felső Duna szakasz diatóma közösségeit és a Velencei tavi, különböző aljzaton előforduló perifitonjának sajátosságait.

Folyóvízi fitoplakton közösségek

Use of Self-Organizing Maps (SOM) for characterisation of riverine phytoplankton associations in Hungary, (Várbíró et al., 2007) 1.

A célkitűzésünk az volt, hogy a magyarországi folyóvizek jellemző fitoplakton közösségeit a meghatározzuk és jellemezzük. Az adatsor a magyarországi monitoring hálózat adatait tartalmazta és ezen adatok adatbázisba történő rendezése után vált értékelhetővé.

Az elemzésben használt adatbázis 1897 mintavétel adatait tartalmazta 189 mintavételi hely esetében. A SOM módszer segítségével nyolc folyóvíztípust lehetett elkülöníteni. Minden típusra meghatároztuk a domináns fitoplakton közösséget és jellemeztük az egyes csoportokat. A fitoplakton taxonokat a Reynolds féle kodonbesorolás alapján elemeztük (Reynolds, 2002). Néhány csoport szoros összefüggést mutatott a környezeti változókkal és lehetővé vált azon folyószakaszok illetve fitoplakton közösségek meghatározása, amelyek nem megfelelő ökológiai állapotot jeleznek.

Az általunk meghatározott folyóvízi fitoplakton típusok (2. ábra):

I. típus A folyók felső szakaszaira legjellemzőbb típus a TIB kodon (bentikus kovaalgák), gyakran 100% abundanciaértékkel. Leggyakoribb taxonok a *Nitzschia palea*, *Nitzschia fonticola*, *Navicula capitatoradiata*, *Surirella brebissonii*, *Diatoma vulgare*.

II. típus Ebben a típusban is a TIB csoport dominál, de jelentős arányban jelen vannak a szerves szennyezésre utaló W0 és W1 funkcionális csoportok fajai (*Chlamydomonas reinhardtii*, *Euglena spp.*).

III. típus Ez a típus a kisebb folyókra jellemző, ahol a víz tartózkodási ideje, hidromorfológiai beavatkozások miatt megnő és lehetővé teszi valódi plankton kialakulását. Jellemző kodonok a TIB és a D csoportok.

IV. típus Ez a típus a Dunára jellemző plaktontípus. Jellemző kodonok a D, J és a C. Domináns fajok a D kodonba tartozó *Stephanodiscus hantzschii*, *Cyclotella comta* *Skeletonema potamos* és *Nitzschia acicularis*, a C kodonba tartozó *Cyclotella meneghiniana* és a J kodonba tartozó *Scenedesmus* fajok.

V. típus Ebbe a típusba a Tisza alsó szakasza tartozik. A domináns kodon az Y, jellemző fajok a *Cryptomonas reflexa*, *C. marssonii*, *C. rostratiformis*.

VI. típus Ebbe a típusba a szerves szennyezéssel terhelt vizek tartoznak. A terhelést jellző W0 kodon dominál, jellemző fajok *Chlamydomonas* spp., *Euglena* spp., (*Polytoma uvella*, *Spermatozopsis exultans*.)

VII. típus A lassú folyású kis csatornák és erek jellemző asszociációja. Több relatívan ritka kodon fordul elő benne, például L0, H1, L, S. Szezonálisan bimodális eloszlás jellemzi, azaz a nyári és a téli vegetációs periódusban is dominánsá válhatnak a H1 másrészt az L0 (Dinofita) funkcionális csoportok fajai.

VIII. típus A W1 kodon dominanciája jellemzi ezt a típust, amely Reynolds szerint a kis szerves tavak jellemző planktonja. Gyakori taxonok a metafitikus *Phacus* és *Trachelomonas* fajok. Ez a típus a dús növényzettel benőtt csatornákra jellemző, tipikusan a nyári vegetációs periódusban előforduló asszociáció.

A Duna bentikus kovaalga flórája

Use of Kohonen Self Organizing Maps (SOM) for the characterisation of benthic diatom associations of the River Danube and its tributaries. (Várbíró et al., 2007)

Ebben a cikkben a Duna felső szakasza, bentikus kovaalga flórájának bemutatása történt meg.

A mintavételi helyek a Duna forrásától a Magyarországi szakaszának végig terjedtek és három éves periódust öleltek fel. A jellemző kovaalga közösségeket SOM módszer segítségével határoztuk meg. Az elkészült SOM térképet *k*-átlag klaszterezési eljárással 5 klaszterre különítettük el. Néhány kivételtől eltekintve az

első három klasztercsoportba a Német-Osztrák Duna-szakasz került, a 4. és 5. csoportok a Szlovák-Magyar szakaszra estek. A mintavételek időbeni különbsége nem volt számottevő. A klaszterek a mintavételi helyek szerint alakultak ki. Az 1-2. klaszter karakterfajai, melyet a strukturális index segítségével határoztunk meg: *Achnanthydium minutissimum*, *Denticula tenuis*, *Nitzschia fonticola*, *Psammodithium bioretii*, *Melosira varians*, *Planothidium subatomoides*. A 3. klaszter karakterfaja az *Amphora pediculus*, míg a 4-5. klaszteré a *Nitzschia dissipata*, *Navicula cryptotenella*; *Cocconeis placentula* var. *euglypta*, *Rhoicosphenia abbreviata*, *Nitzschia inconspicua* voltak.

Aljzat szerepe kovaalga közösségek kialakulásában

Comparative algological and bacteriological examinations on biofilms developed on different substrata in a shallow soda lake (Ács et al., 2008)

A felmérés a Velencei tó gazdag perifiton közössének vizsgálatára irányult. Arra a kérdésre kerestük a választ, hogy meghatározzuk az eltérő aljzatokon előforduló kovalaga közösségek összetételét.

A SOM analízis alapján a kovalga közösségek abundancia viszonyai a májusi mintavételek esetén két csoportra különíthetők el. Az avas nád, az andezit és a gránit aljzatok alkottak egy csoportot, a zöld nád és a polikarbonát pedig egy másikat. Az *Achnanthydium minutissimum* és *Navicula cryptotenella*, mint pionír kovaalga kolonista fajok, erőteljesen meghatározók voltak valamennyi aljzat esetén.

Novemberben a SOM analízis alapján az avas és a zöld nád az andezit és gránit aljzattal alkotott egy csoportot, míg a polikarbonát a plexi aljzattal. Ez a hasonlósága az *Achnanthydium minutissimum*, fajnak volt köszönhető, ami az utóbbi két szubsztráton dominált. További jellemző fajok voltak még a *Fragilaria pulchella* és a *Nitzschia perminuta*. A zöld és avas nád jellemző bevonatalgái a SOM komponens térképek alapján a *Nitzschia palea* és a *Gomphonema olivaceum* var. *olivaceum* voltak.

Folyóvizek fitoplakton alapú minősítése

A new evaluation technique of potamoplankton for the assessment of the ecological status of rivers. (Borics et al., 2007)

A folyóvizek fitoplaktonjának monitorozása során a természetes algaközösségek fajszerkezetéről, fajösszetételéről, diverzitásáról nyerünk információt. Azon szakemberektől akik, a vízminőség védelem területén dolgoznak nem várható el, hogy tudományos részletességű eredményekből és fajlistákból hámozzák ki a megfelelő információt. Számukra jól értelmezhető, egyszerű indexeket kell kidolgozni.

Egészen napjainkig kevés sikeres próbálkozás irányult arra, hogy a folyóvizek ökológiai állapotát a fitoplankton alapján értékeljük. Jelen tanulmányukban egy olyan értékelő rendszert mutatunk be, amely a folyóvizek algaflóráján alapul.

A módszer az eredetileg állóvizekre kidolgozott fitoplakton funkcionális csoportokon alapul (Reynolds, 2002). Az értékelés eredményeként egyetlen számmal, Q index, jellemezzük a vizek ökológiai állapotát. Az indexet különböző magyarországi víztesten teszteltük és jóval érzékenyebbnek bizonyult, mint a korábban használt szaporítási index.

Magyarországi felszíni vizei ökológiai állapotának áttekintése bentikus kovaalga közösségek alapján

*Implementation of the European Water Framework Directive:
Development of a system for water quality assessment of Hungarian
running waters with diatoms (Van Dam et al., 2007)*

Ebben a cikkben az első országos kovalga felmérés eredményei kerültek bemutatásra. Az adatsor megfelelően robusztus volt, mert mind a mintavételi, mind a mintafeldolgozási, határozási és számolási hibákat a lehető legminimálisra csökkentettük.

Ezen felül a diatoma vizsgálatokkal párhuzamosan más, a Vízkeretirányelv hatálya alá tartozó biológiai minőségi elem (halak, makroszkópikus vízi gerinctelenek, fitoplakton, makrofita) vizsgálatára is sor került, valamint kémiai és hidromorfológiai felmérés is készült. Közel 359 kovalga taxon került elő a vizsgált 339 mintavételi helyről. A leggyakoribb genus a *Navicula* (104 db faj) volt, amit a követett a *Nitzschia* (80), *Fragilaria* (43), *Achnanthes* (41), *Gomphonema* (30) és *Cymbella* (26). A leggyakrabban előforduló taxon az *Achnanthes minutisimum* volt, amely a leggyakoribb diatoma taxon a világon. A legtöbb előkerült taxon a meszes, eutróf vizekre jellemző. Számos a szennyezésre

toleráns faj is előfordult (pl. *Gomphonema parvulum*, *Nitzschia paleacea*), de a szennyezéstől mentes vizeket jelző fajok is előkerültek (pl. *Achnanthes biasolettiana*, *Gomphonema micropus*). A *Meridion circulare* a kevés tipikusan folyóvízre jellemző kovaalga egyike, szintén jellemző volt a mintákban (Van Dam et al., 1994). A savas és oligotróf vizekre jellemző fajok ritkán került elő pl. *Eunotia bilunaris*. Az adatelemzésben felhasznált kovalaga adatbázis az ECOSURV project adatbázisának része.

Ezen adatbázis képezi az alapját a jelenleg az országos monitoring hálózat által használt minősítő és adattároló rendszereknek.

NEURAL NETWORKS IN ECOLOGICAL DATA ANALYSIS

Introduction and objectives

The need for better techniques, tools and practices to analyse ecological and economic systems within an integrated framework at wider scales has never been so great.

Over the last few years, the use of Artificial Neural Network (ANN) techniques has provided ecologists with a powerful method for modelling large volumes of ecological data (Paisley et al., 2003). Neural networks are powerful computational tools that can be used for classification, pattern recognition, empirical modelling and for many other tasks.

Ecosystem analysis and prediction with empirical statistical and analytical methods are often limited by the spatial complexity and temporal dynamic character of ecological processes. This is one reason for the typically non-linear interrelations of variables and species with data being not normally distributed. Artificial neural networks provide an attractive alternative tool for analysing ecological data and for modelling due to their specific features such as non-linearity, adaptivity, generalisation and model independence (no a-priori model needed). ANNs have been applied to various fields of aquatic sciences and engineering, such as modelling water quality (e.g. Daniell and Wundke, 1993; Lachtermacher and Fuller, 1994; Schizas et al., 1994; Kaluli et al., 1998; Wen and Lee, 1998) and relating community characteristics with environmental variables (e.g. Chon et al., 1996; Lek et al., 1996; Recknagel et al., 1998; Guegan et al., 1998; Lee et al., 1998; Maier et al., 1998).

The biological monitoring of the water currents in several countries provided large volume datasets during the last decades. In case of large volume of data, the traditional multivariate statistical methods like cluster analysis and ordination is difficult to interpret and present the information. Self Organizing Map is a novel approach for the visualization of high-dimensional data. Self-organizing map is a data visualization technique invented by Professor Teuvo Kohonen which reduce the dimensions of data through the use of self-organizing neural networks. SOM converts complex statistical

relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. Thus, it compresses information while preserving the most important topological and metric relationships of the primary data items (Kohonen 2001). The main advantages of the SOM are the better data visualisation and noise reduction (Vesanto and Alhoniemi 2000).

Material and Methods

For statistical analysis, we used the SOM algorithm which can make a projection of the database into a two dimensional hexagonal map. Closely related communities were placed into neighbouring hexagons by their similarities, while samples with different communities were in distant hexagons. The SOM can display the groupings of samples and species together; therefore, each species can be evaluated by its importance (Kohonen, 2001). The SOM Toolbox (<http://www.cis.hut.fi/projects/somtoolbox>) was used to implement the SOM under a MATLAB™s environment. Input layer gets information from data matrix, while output layer visualizes the computational results. When an input vector is sent through the network, each neuron of the network computes the distance between the weight vector and the input vector. The output layer consists of output neurons, which are usually arranged into a two-dimensional grid for better visualization. The best arrangement for the output layer is a hexagonal lattice, as it does not favour horizontal and vertical directions as much as a rectangular array. (Fig. 1.)

The algorithm of the SOM process can be described as:

- Step 1: Epoch $t=0$, the virtual units (VU_p) are initialized, weights drawn from the input dataset.
- Step 2: A sample unit (SU_j) is randomly chosen as an input unit.
- Step 3: Using some metric, the distances between SU and each virtual unit are computed.
- Step 4: The virtual unit VU_c closest to the input SU_j is chosen as the winning neuron. VU_c is called the Best Matching Unit (BMU).

- Step 5: The virtual units VU_p are updated with the learning rule
- Step 6: Increase time t to $t+1$. If $t < t_{\max}$ then go to step 2 else stop the training.

In step 1, the user decides the size of the map, the optimal numbers of map nodes according to Vesanto (Vesanto, 2000) can be calculated as (eq. 1.):

Where Nn is the number of nodes in the final map and SU_d is the number of sample units.

In step 3, the distance or the measure of dissimilarity between stations can be chosen. Usually the conventional Euclidean distance is used. (eq. 1.)

In step 5, in the Euclidean learning rule (eq. 3.) the function $h_{ck}(t)$ is the neighbourhood function. During the learning process, the BMU defined in step 4 is not the only updated unit. In the grid, a neighbourhood is defined around the BMU and all units within this neighbourhood are also updated. The most often used function to select the neighbourhood is the Gaussian function (eq. 4.):

$\| \mathbf{r}_k - \mathbf{r}_c \|$ is the Euclidean distance on the map between the winning unit VU_c and each virtual unit VU_k . σ is a decreasing function of time, which defines the width of the part of the map, affected by the learning. α is the 'learning-rate factor'.

Both factors are decreasing functions of the time and converge towards 0.

In order to characterise the VU groups we used the Structuring Index (SI) which was originally developed to define the species that show the strongest influence on the organization of the SOM map (Park et al., 2005). Taxa showing strong gradient display have high SI values, whereas species showing weak gradient present low SI values. Thus, the higher the value of SI, the more relevant the variable is to the structure of the map. The equation calculates the SI value (eq. 5.), where w_{ij} and w_{ik} are the connection weights of the i^{th} species in the SOM units, $\|r_j - r_k\|$ is the topological distance between units j and k , S is the total number of SOM output unit.

The map obtained after the learning process of the SOM represents all the SU assigned to neurons so that similar samples are located close to each other and far from those dissimilar. However, it is not easy to distinguish subsets because there are still no boundaries between possible clusters.

For clustering the SOM we can use several conventional clustering techniques however in our work we used the K-means clustering technique.

For determining the suitable cluster size we used Davies-Bouldin Validity Index (Davies and Bouldin, 1979)

An Ecological database is a structured collection of records or data. The most commonly used database model today is the relational model. The basic data structure of the relational model is the table, where information is represented in columns and rows. Thus, the "relation" in "relational database" refers to the various tables in the database; a relation is a set of tables. The columns can be expressed as various attributes of the entity (sampling date, site name etc.), and a row is an actual instance of the entity (like 2009.05.27., Concó, Ács). One of the strengths of the relational model is that any value occurring in two different records implies a relationship among those two records. It also minimizes the redundant information

Results

The thesis contains five scientific articles. In the theses I described in details the SOM algorithm as well as the basic structure of an ecological database. By combining the database structures and the SOM algorithm we were able to characterize frequent Hungarian running water types according to their characteristic phytoplankton assemblages.

I constructed the phytoplankton and phytobenton database which is presently in use in the environmental authorities, and is concerned as National Database.

By the application of SOM methods we characterized the upper Danube section based on their characteristic diatom communities, and described the differences and similarities of the periphyton of the Lake Velencei on different substrates.

Characteristic riverine phytoplankton associations

Use of Self-Organizing Maps (SOM) for characterisation of riverine phytoplankton associations in Hungary, (Várbiro et al., 2007)

The phytoplankton database of the Middle Danube Basin was analysed and evaluated in order to describe the characteristics algal assemblages of the rivers. The dataset were extracted from the database of the Hungarian monitoring network and academic institutions. The database which have been included in the analysis contains the results of 1897 phytoplankton investigation from 189 locations. Implementing the SOM method we can identify the different algal communities which characterise different river types. The algal communities were described as different ratios of algal functional groups. Since some of the groups are in close relation with certain types of environmental pressure it is also possible to highlight those rivers or river sections (or those periods) which are far from the expected good ecological status. We identified eight clusters or river types and described their composition.

Type I. This is the most frequent phytoplankton type which is representative for the upper section of rivers. The dominant codon is the TIB (bentic diatoms). This association has no seasonal preference. The most frequent taxa are *Nitzschia palea*, *Nitzschia fonticola*, *Navicula capitatoradiata*, *Surirella brebissonii*, *Diatoma vulgare*.

Type II. Members of the TIB group are dominating type but the elements of the Wo and W1 functional groups (*Chlamydomonas reinhardtii*, *Euglena spp.*) indicate organic pollution.

Type III. This is typical plankton of small rivers with the dominance of TIB and D codons. This assemblage occurs on those river sections where the retention time due to hydro-morphological changes (upper stream reservoirs) enables real phytoplankton to develop.

Type IV. This association can be referred to as "Danube type" summer plankton. The most important codons are D, J and C. Dominant species of this association are *Stephanodiscus hantzschii*, *Cyclotella comta* *Skeletonema potamos* and *Nitzschia acicularis* from codon D, *Cyclotella meneghiniana* from codon C, and *Scenedesmus* species from codon J.

Type V. This type could be mainly found on the lower section of the river Tisza with the dominance of Y codon. *Cryptomonas reflexa*, *C. marssonii*, *C. rostratiformis* become dominant in this section.

Type VI. The characteristic functional group of type VI is Wo. The dominance of this group is due to very strong organic pollution. *Chlamydomonas* spp., *Euglena viridis*, *Polytoma uvella*, *Spermatozopsis exultans* are typical elements of this assemblage. This type of plankton is usually dominant in winter and early spring.

Type VII. This type is a mixture of very diverse associations with the presence of relatively rare groups like Lo, HI, L, S. The occurrence of this type is expected in slow flowing channels and small rivulets in late winter and summer. The bimodal character of the type VII is caused by the dinophytes (Lo functional group.)

Type VIII. The dominant codon of this type is the W1 which is according to Reynolds characteristic for small organic pools. Frequent taxa are the metaphytic *Phacus* and *Trachelomonas* spp. This type is typical in slow-flowing rivulets and channels which are under the risk of organic pollution and have rich macrophyte vegetation. Development of this type is expected in summer.

Benthic diatome associations of the Danube

Use of Kohonen Self Organizing Maps (SOM) for the characterisation of benthic diatom associations of the River Danube and its tributaries. (Várbíró et al., 2007)

In this paper the investigation of the benthic diatom flora of the River Danube is presented. The investigations include samples from the source streams to the end of the Hungarian stretch during a three years period. Characteristic diatom assemblages were established using SOM algorithm. K-means clustering of the SOM allocated 5 different sample groups. With few exemptions, the first, second and third clusters fell in the German-Austrian stretch, whereas the fourth and fifth clusters fell in the Slovakian-Hungarian stretch. Temporal differences were less pronounced: the five clusters were mainly organized by the sampling localities. We found that the clusters 1 and 2 were characterized by -according to their high structuring index- *Achnanthes minutissimum*, *Denticula tenuis*, *Nitzschia fonticola*, *Psammodictyon bioretii*, *Melosira varians*, *Planorbulina subatomoides*. The character species of cluster 3 is *Amphora pediculus*, while character species of the clusters 4 and 5 are *Nitzschia dissipata*, *Navicula cryptotenella*; *Cocconeis placentula* var. *euglypta*, *Rhoicosphenia abbreviata*, *Nitzschia inconspicua*.

Periphyton communities on different substrata

Comparative algological and bacteriological examinations on biofilms developed on different substrata in a shallow soda lake (Ács et al., 2008)

The periphyton of Lake Velencei which is a diverse and dense community was investigated in this article. The main interest was to distinguish the differences of periphyton communities according to the different substrate type.

According to the SOM analysis based on algological data (species composition and their abundance), samples of different substrata formed two distinctive groups. Old reed, andesite and granite were grouped together, respectively, and polycarbonate and green reed clearly separated from the other three in May. *Achnanthes minutissimum* and *Navicula cryptotenella*, a pioneer

colonist diatom species were found to be vigorously dominant on all substrata.

In November, based on the SOM analysis, granite and andesite come together with green and old reeds comprised one group, plexi and polycarbonate the other. These similarity was due to the presence of *Achnantheidium minutissimum*, strongly dominating both plexi and polycarbonate substrata. Characteristic species were also *Fragilaria pulchella* and *Nitzschia perminuta* for these type of substrates, while for green and old reed a clear gradient of increasing of *Nitzschia palea* can and *Gomphonema olivaceum* var. *olivaceum* be viewed from the SOM component planes.

Phytoplankton index development in rivers

A new evaluation technique of potamoplankton for the assessment of the ecological status of rivers.(Borics et al., 2007)

Monitoring of the naturally occurring algal communities in rivers provide data on species composition, species number, diversity, or quantitative occurrence of the algae. Experts of administrative institutions who are responsible for water quality management need simple numerical values rather than species lists or scientific evaluation of the assemblages. Since up to now, little attention has been attended to the application of the phytoplankton for evaluation of ecological state of the rivers, the present study aims to show a new assessment method. The method is based on the functional group of algae, described for lakes by Reynolds (Reynolds, 2002), represented in the potamoplankton and provides a single index number (Q). The index has been tested on phytoplankton data of different rivers, and proved to be more sensitive than the earlier used saprobic index. The authors task in this project was to build the database and make the calculation algorithm for the dataset.

Nation wide survey of Hungary by diatomes

*Implementation of the European Water Framework Directive:
Development of a system for water quality assessment of Hungarian
running waters with diatoms* (Van Dam et al., 2007)

In this study the results of the first nation wide survey of benthic river diatoms is presented. The dataset is robust as the primary sources of uncertainty in diatom training sets like choice of site and substrate for sampling, inter-operator differences in diatom taxonomy and counting techniques were eliminated as much as possible. Moreover, parallel with the diatoms, investigations were made on other biological quality elements, including macro-invertebrates and fish. In total, nearly 500 taxa were found in the 339 sites. The commonest genera are *Navicula* (104 taxa), followed by *Nitzschia* (go), *Fragilaria* (43), *Achnanthes* (41), *Gomphonema* (30) and *Cymbella* (26). The most common taxa are listed in Table 2. The most common species, *Achnantheidium minutissimum*, is also the most common freshwater diatom species in the world. The majority of the other common taxa are characteristic of eutrophic, alkaline waters. Some of these are tolerant to organic pollution (e. g. *Gomphonema parvulum*, *Nitzschia paleacea*), others require cleaner waters (e. g. *Achnanthes biasolettiana*, *Gomphonema micropus*). *Meridion circulare* is one of the very few diatom species which are characteristic for running waters (Van Dam et al., 1994). Species from acidic or oligotrophic waters are rare in the dataset: the most common one is *Eunotia bilunaris*. The diatom database used in the paper was part of the ECOSURV database. The author of the theses was responsible for building the database structure as well as providing the relevant data matrixes.

This database is the basic of the recently used in the Hungarian monitoring network for quality assessment and data storage.

Irodalom / References

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.

Daniell, T.M., Wundke, A.D. 1993. Neural networks—assisting in water quality modelling. Preprints, Watercomp, Melbourne, Australia, pp. 51–57.

Davies, D.L., Bouldin, D.W. 1979. A cluster separation measure. 1979. *IEEE Trans. Pattern Anal. Machine Intell.* 1 (4). 224–227.

Guegan, J.-F., Lek, S., Oberdorff, T. 1998. Energy availability and habitat heterogeneity global riverine fish diversity. *Nature* 391, 382–384.

Kaluli, J.W., Madramootoo, C.A., Djebbar, Y. 1998. Modelling nitrate leaching using neural networks. *Water Sci. Technol.* 38 (7), 127–134.

Kohonen, T. 2001. *Self-organizing Maps*. 3rd ed. - Springer, Berlin

Lee, H.-L., DeAngelis, D., Koh, H.-L. 1998. Modelling spatial distribution of the unionid mussels and the core-satellite hypothesis. *Water Sci. Technol.* 38 (7), 73–79.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M. 1996. Role of some environmental variables in trout abundance models using neural networks. *Aquat. Living Resour.* 9, 23–29.

Maier, H.R., Dandy, G.C., Burch, M.D. 1998. Use of artificial neural networks for modelling cyanobacteria *Anabena* spp. in the River Murray, South Australia. *Ecol. Model.* 105, 257–272.

Paisley, M.F., Walley, W.J., Nikhade, J., Dils, R. 2003. Identification of the key biological indicators of nutrient enrichment in rivers for use in predictive/diagnostic models. *Diffuse Pollution Conference, Dublin 2003 8A Ecology*.

Park, Y.-S., Gevrey, M., Lek, S., Girard, J. 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., Verdonscho, P., Park, Y. S., Descy, J. P. (eds.): *Modelling Community Structure in Freshwater Ecosystems*: 369-380. Springer, Berlin.

Recknagel, F., Fukushima, T., Hanazato, T., Takamura, N., Wilson, H. 1998. Modelling and prediction of phyto- and zooplankton dynamics in Lake Kasumigaura by artificial neural networks. *Lakes Reserv. Res. Manag.* 3, 123–133.

Reynolds, C., Huszar, V., Kruk, C., Nasseli-Florel, S., Melo, S. 2002. Towards functional classification of the freshwater phytoplankton. - *J. Plankton Res.* 24: 417- 428.

Schizas, C.N., Pattiches, C.S., Michaelides, S.C. 1994. Forecasting minimum temperature with short time-length data using artificial neural networks. *Neural Netw. World* 4 (2), 219–230.

Van Dam, H., Mertens, A., Sinkeldam, J. 1994. A coded checklist and ecological indicator values of freshwater diatoms from The Netherlands. - *Netherl. J. Aquat. Ecol.* 28: 117-131.

Vesanto, J. 2000. Neural network tool for data mining: SOM Toolbox. - *Proc. Symp. On Tool Environments and Development*

Methods for Intelligent Systems (TOOLMET2000): 184.196. Oulun yliopistopaino, Oulu, Finland.

Vesanto, J., Alhoniemi, I. 2000. Clustering of the self-organizing map. - IEEE Transact. on Neural Networks 11: 586-600.

Wen, C.-G., Lee, C.-S. 1998. A neural network approach to multiobjective optimization for water quality management in a river basin. Water Resour. Res. 34, 427-436.

Várbíró Gábor tudományos tevékenységének jegyzéke

1.1 Témával kapcsolatos cikkek

1. **Várbíró, G.**, Borics, G., Kiss, K.T., Szabó, K. É., Ács, É. (2007): Use of Kohonen Self Organizing Maps (SOM) for the characterisation of benthic diatom associations of the River Danube and its tributaries. – Arch. Hydrobiol. Suppl. Large Rivers 17: 395-403. **IF: 1.409**
2. **Várbíró, G.**, Ács, É., Borics, G., Érces, K., Fehér, G., Grigorszky, I., Japport, T., Kocsis, G., Krasznai, E., Nagy, K., Nagy-László, Zs., Pilinszky, Zs., Kiss, K.T. (2007): Use of Self-Organizing Maps (SOM) for characterisation of riverine phytoplankton associations in Hungary. - Arch. Hydrobiol. Suppl. Large Rivers 17: 383-394. **IF: 1.409**
3. Borics, G, **Várbíró, G.**, Grigorszky, I, Krasznai, E., Szabó, S., Kiss, K. T. (2007): A new evaluation technique of potamoplankton for the assessment of the ecological status of rivers. Arch. Hydrobiol. Suppl. Large Rivers 17: (3-4) 465-486 . **IF: 1.409**
4. Ács, É., Borsodi, A.K., Kiss, É., Kiss, K.T., Szabó, K.É., Vladár, P., **Várbíró, G.**, Záray, Gy. (2008): Comparative algological and bacteriological examinations on biofilms developed on different substrata in a shallow soda lake – Aquat Ecol (2008) 42:521–531. **IF: 0.925**
5. Van Dam, H., Stenger-Kovács, Cs., Ács, É., Borics, G., Buczkó, K., Hajnal, É., Soróczki-Pintér, É., **Várbíró, G.**, Tóthmérész, B., Padisák, J. (2007): Implementation of the European Water Framework Directive: Development of a system for water quality assessment of Hungarian running waters with diatoms. - Arch. Hydrobiol. Suppl. Large Rivers 17: 339-364. **IF: 1.409**

1.2 Egyéb tudományos közlemények

6. Borics, G., Tóthmérész, B., Grigorszky, I., Padisák, J., **Várbíró, G.**, Szabó, S. (2003): Algal assemblage types of bog-lakes in

Hungary and their relation to chemistry, hydrological conditions and habitat diversity. *Hydrobiologia* 502: 145-155. **IF: 0.720**

7. Krasznai, E., Fehér, G., Borics, G., **Várbíró, G.**, Grigorszky, I. & Tóthmérész, B. (2008): Use of desmids to assess the natural conservation value of a Hungarian oxbow (Malom-Tisza, NE-Hungary). *Biologia Bratislava* 63/6: 924—931, **IF: 0,207**
8. Ács, É., Borics, G., Fehér, G., Kiss, K.T., Reskóné, N. M., Stenger-Kovács, Cs., **Várbíró, G.** (2009): Implementation of the European Water Framework Directive to assess the water quality in Hungarian running waters with diatoms. – *Diatomededelingen* 33: 29-32. ISSN1872-9673
9. Borics, G., Krasznai, E., **Várbíró, G.**, Abonyi, A., Grigorszky, I., Szabó, S. (2008): Néhány Tisza-menti holtág jellegzetes fitoplankton asszociációi. *Hidrológiai Közlöny* 88: (6) 34-36
10. Grigorszky, I., Borics G., **Várbíró, G.**, Grigorszky, T., Béres, V., Krasznai, E., Schnitcher, Cs., (2008): A nitrát és nitrit ion hatása az üledékből történő foszfor felszabadulásra két Körös-vidéki holtmederben *Hidrológiai Közlöny* 88: (6) 61-63
11. Ács, É., Borics, G., Kiss, G., Reskóné, Nagy, M., **Várbíró, G.**, Kiss, K.T. (2007): Fitobenton vizsgálatok tanulságai a Balaton és vízgyűjtője patakjainak példáján. – MHT XXV. Országos Vándorgyűlés. Tata, 2007. július 4-5. pp. 1-27. CD. - ISBN 978-963-8172-20-4

1.3 Előadások poszterek

12. Ács, É., Bolla, B., Kiss, K.T., Reskóné Nagy, M., **Várbíró, G.** (2009): A bevonatalkó kovaalgák szerepe a Balaton ökológiai állapotának megítélésében. – leadott absztrakt a 8. Magyar Ökológus Kongresszusra.
13. Krasznai, E., Borics, G., **Várbíró, G.**, Abonyi, A., Padisák, J., Tóthmérész, B. (2008): Spatial and temporal pattern of vertical

distribution of phytoplankton in a shallow oxbow (Hungary). – The 15th workshop of the International Association of phytoplankton taxonomy and ecology (IAP), Israel

14. Borics, G., **Várbíró, G.**, Krasznai, E., Grigorszky, I. (2008): Milyen mértékben határozza meg a meder morfológiája sekély tavaink trofitását? – L. Hidrobiológus Napok, Hungary,
15. Krasznai, E., Borics, G., **Várbíró, G.**, Tóthmérész, B. (2008): Cianobaktérium dinamika egy sekély holtágban. – L. Hidrobiológus Napok, Hungary,
16. Krasznai, E., Fehér, G., Borics, G., **Várbíró, G.**, Grigorszky, I. & Tóthmérész, B. (2007): Use of desmids in the assessment of the natural value of a Hungarian oxbow – The International Symposium on Biology and taxonomy of green algae V. Smolenice Castle, June 25-29, 2007, Slovakia, Programme & Abstracts
17. Ács, É., Szabó, K.É., **Várbíró, G.**, Kiss, K.T., Borics G. (2007): Use of Kohonen Self Organizing Maps (SOM) for the characterisation of benthic diatom associations of the River Danube and its tributaries. 6th International Symposium Use of Algae for Monitoring Rivers p. 20.
18. Ács, É., Borics, G., Kiss, G., Reskóné, Nagy, M., **Várbíró, G.**, Kiss, K.T. (2007): Fitobenton vizsgálatok tanulságai a Balaton és vízgyűjtője patakjainak példáján. – MHT XXV. Országos Vándorgyűlés. Tata, 2007. július 4-5. pp. 1-27. <http://www.hidrologia.hu/ovgytata/25/7szekcio/250700.htm>
19. Borics, G., Krasznai, E., Abonyi, A., **Várbíró, G.**, Grigorszky, I. (2007): Néhány Tisza menti holtág jellegzetes fitoplankton asszociációi – IL. Hidrobiológus Napok, 2007, Hungary,
20. Grigorszky, I., Borics, G., **Várbíró, G.**, Grigorszky, T., Béres, V., Krasznai, E., Schnitcher, Cs. (2007): A nitrát és nitrit hatása a holtágak üledékéből történő foszfor pótlására – IL. Hidrobiológus Napok, Hungary,
21. Borics, G., **Várbíró, G.**, Kiss, K. T., Grigorszky, I., Krasznai, E. (2006): Possible evaluation of the rheo-plankton for assessing th

ecological status of rivers. 6th International Symposium Use of Algae for Monitoring Rivers p. 29.

22. **Várbíró, G.**, Ács, É., Borics, G., Ércses, K., Fehér, G., Grigorszky, I., Japport, T., Kocsis, G., Krasznai, E., Nagy, K., Nagy-László, Zs., Pilinszky, Zs., Kiss, K. T. (2006): Use of Self Organizing Maps (SOM) for characterisation of riverine phytoplankton associations in Hungary. 6th International Symposium Use of Algae for Monitoring Rivers p. 172.
23. Grigorszky, I., Kiss, K. T., Dévai Gy., Nagy, A.S., Borics, G., Gligora, M., Kraj K., Plenkovich, A., Béres, V., Owsianny, P., **Várbíró G.** (2006): The main charecteristics of phytoplankton populations in the hungarian section of the River Tisza in connection with the WFD. 6th International Symposium Use of Algae for Monitoring Rivers p. 53.
24. Krasznai, E., **Várbíró, G.**, Borics, G., Grigorszky, I. Some interesting algological characteristics of the Tiszadob-oxbow. 16th Hungarian Algological Meeting 2005. Abstract book. 19.
25. **Várbíró, G.**, Borics, G. (2004): Establishment of the East Hungarian Biomonitoring Network (EHBN), In Klumpp, A., Ansel, W., Klumpp, G., (Eds,) Urban air Pollution Bioindication and environmental awareness. 321-327; Cuvillier Verlag, Göttingen.
26. **Várbíró, G.**, Borics, G. (2002): Establishment of the East Hungarian Biomonitoring Network (EHBN), 2002. Oradea
27. **Várbíró, G.**, Borics, G. (2001): The establishment of the East Hungarian Biomonitoring Network. EURASAP 41. Newsletter. April 2001. p.36-37. ISSN-1026-2172. Sofia.
28. Borics, G., Grigorszky, I., Padisák, J. and **Varbíró, G.** (2001): Two morphologically and ecologically different types of *Peridinium bipes* Stein in Hungarian waters. Plankton symposium 2001. Abstract book. p. 222.
29. **Várbíró, G.**, O'Heix, B.-C., Tóth, J. A., Dizengremel, P., Garrec, J.P., (1998): Air quality mapping of the city Debrecen with epiphytic green algae as biomonitors.. Microflora et Pollution. Egregia, Nyiregyhaza-Nancy, pp. 93-102.

30. O'Heix, B.-C., **Várbíró, G.**, Tóth, J. A., Dizengremel, P., Garrec, J.P., Balazsy S. (1997): Use of epiphytic green algae growing on tree trunks as biomonitors of air quality. International Regional Seminar Environmental Protection: Modern Studies in Ecology and Microbiology. Ukraine Uzhgorod., Vol., 1. pp. 32-39.
31. O'Heix, B.-C., **Várbíró, G.**, Tóth, J. A., Dizengremel, P., Garrec, J.P. (1997): Method for the air quality mapping of a city thanks the use of biomonitoring example of Debrecen - Hungary.. International Regional Seminar Environmental Protection: Modern Studies in Ecology and Microbiology. Ukraine Uzhgorod, Vol., 2. pp. 50-55.