INFLUENCE OF CLIMATE CHANGE ON DIATOMS DIVERSITY INDICES IN LAKE PRESPA

Andreja Naumoski
Kosta Mitreski

1. INTRODUCTION

Applying machine learning techniques into ecology have proven to be useful into obtaining knowledge for certain problems. Using these diversity indices (DIs) it will be very useful to model the specific diatom communities which are known to exist only in definite environmental conditions. This property is used to model the abiotic environment influence on diatoms.

Using the diatom relative abundances into account, a diversity index depends not only on species richness but also on the evenness, or equitability, with which individuals are distributed among the different species. Diversity indices provide important information about rarity and commonness of species in a community. The ability to quantify diversity in this way is an important tool for biologists trying to understand community structure (River diatoms: a multiaccess key). Understanding how these indexes interact with physical-chemical parameters of the given environment is very useful to know. Parameters like temperature, dissolved oxygen, pH, ammonia and others are one of the few that are vital for diatom survival (Reynolds. C. S., 1998). This is why we build models to see how diatoms diversity indices response on the changes of these parameters.

In order to extract this knowledge from the ecological data we use machine learning techniques. The most researched type of machine learning is inductive machine learning, where the experience is given in the form of learning examples. Machine learning (and in particular predictive modelling) is increasingly often used to automate the construction of ecological models (Džeroski, 2001), (Joergensen, 2001). Most frequently, models with regression trees of diversity indices and population dynamics are constructed from measured data by using machine learning techniques. The most popular machine learning techniques used for modelling diversity indices include decision tree induction and rule induction.

In this paper, we focus on applications of machine learning in ecological modelling, more specifically, applications of modelling diversity indices. We will use a dataset, which has been collected from different measurement stations placed in Lake Prespa, as a part of the EU project TRABOREMA (TRABOREMA Team, 2005-2007). Several important parameters are measured, which reflect the physical, chemical and biological aspects of the water quality of the lake. From these measurements, several diatoms (algae) belonging to the group Bacillariophyta) will be considered for estimating a relationship between their relative abundance, and then calculated their diversity indices and the abiotic characteristics of the habitat. Diatoms are known to be almost ideal bio-indicators of the environment in several studies (Van Dam H., 1994).

The paper is organized as follows. Section 1 introduces with idea of the diversity indices modelling and the main purpose of this paper, the diversity indices models for Lake Prespa, in Section 2 we give an overview of the diversity indices modelling and introduction to machine learning, with a briefly description of the approach to machine learning that is often used in this kind of modelling: decision tree induction and rule induction. The measured data, the main diatom bio-indicators and data collection procedures are presented in Section 3, while Section 4 describes the diversity indices models which were built for several diatoms from lake and rivers measurements. Section 5 concludes and gives directions for future work on this subject.

2. DIVERSITY INDECES MODELLING

The output of a diversity indices model is some property of the population of the target group of organisms at the spatial unit of analysis. There are two degrees of freedom here: one stems from target property, the other from the group of organisms studied. In the simplest case, the output is just the presence/absence of a single species (or group). In this case, we simply talk about diversity indices models.

The input to a DIs model is a set of environmental variables, which in our case are two different kinds. The first kind concerns abiotic properties of the environment, e.g., physical and chemical characteristics. The second kind concerns some biological aspects of the environment, which may be considered as an external impact on the group of organism under study. In our case the biological aspects of the environment are the diversity indices of diatoms abundance.

2.1 Machine learning for diversity indices modelling

The input to a machine learning algorithm is most commonly a single flat table comprising a number of fields (columns) and records (rows). In general, each row represents an object and each column represents a property. In machine learning terminology, rows are called examples and columns are called attributes (or sometimes features). Attributes that have numeric (real) values are called continuous attributes. Attributes that have nominal values are called discrete Attributes. The tasks of classification and regression are the two most commonly addressed tasks in machine learning.

They are concerned with predicting the value of one variable from the values of other variables. The target variable is called the class (dependent variable in statistical terminology). The other variables are called attributes (independent variables in statistical terminology). If the class is continuous, the task at hand is called regression. If the class is discrete (it has a finite set of nominal values), the task at hand is called classification. In both cases, a set of data is taken as input, and a predictive model is generated. This model can then be used to predict values of the class for new data. The common term predictive modelling refers to both classification and regression. Given a set of data (a table), only a part of it is typically used to generate (induce, learn) a predictive model. This part is referred to as the training set. The remaining part is reserved for evaluating the predictive performance of the learned model and is called the testing set. The testing set is used to estimate the performance of the model on new, unseen data.

2. 2 Decision Trees

Decision trees are hierarchical structures, where each internal node contains a test on an attribute, each branch corresponds to an outcome of the test, and each leaf node gives a prediction for the value of the class variable. Depending on whether we are dealing with a classification or a regression problem, the decision tree is called a classification or a regression tree, respectively.

Model trees, where leaf nodes can contain linear models predicting the class value, represent piecewise linear functions. Most algorithms for decision tree induction consider axis-parallel splits. However, there are a few algorithms that consider splits along lines that need not be axis-parallel or even consider splits along non-linear curves. A commonly used procedure for estimating the performance on unseen cases is cross – validation.

2.3 Diversity indices for physical-chemical modelling

Diversity indices provide important information about rarity and commonness of species in a community. The ability to quantify diversity in this way is an important tool for biologists trying to understand community structure. By taking relative abundances into account, a diversity index depends not only on species richness but also on the evenness, or equitability, with which individuals are distributed among the different species.

In this paper we present two indices which have the greatest correlation between the index of the diatoms and the abiotic parameters of the environment: Shannon diversity and Shannon Evenness using the WEKA and CLUS systems for machine learning. These indices are important to estimate dominance of the spices in the community, evenness and richness of the community. In this way, every index is important and depends only of the questions we ask about the species that are enrolled in the experiment. We are interested to model the diatoms community, which latter will be used as bio indicator of the environment.

Nevertheless, we have taken into the experiment several other indices, but most of the testing correlation coefficients are very low. The influences of the abiotic factors on the community structure are captured using the regression trees from both systems.

3. DATA FOR LAKE PRESPA

3.1 Data acquisition methods and instruments

This section provides information about the instruments and procedures used for measuring data. The physical parameters: Temperature, Conductivity, pH, Transparency and Dissolved Oxygen are field measured with HANNA instrument (TRABOREMA Team, 2005-2007). The Oxygen Saturation and the Oxygen Deficiency are obtained by mathematical analysis, while the BOD is obtained by the HANNA instrument. The chemical aspect of water quality is represented by measuring NH4, NO3, NO2, Total_N, Organic_N, Inorganic_N, Total_P and SO4 obtained by a spectroscope from an analytical set. The rest of the chemical elements: K, Na, Mg, Cu, Mn, Zn are measured analytically with AAS by wet digestion, while the chlorophyll data are obtained with an analytical procedure from the spectroscope, following the extraction. Samples for analysis were taken from the surface water of the lake at several locations.

In the usual environmental monitoring and screening (like the one during the TRABOREMA project) diatom cells are collected by a planktonic net or as an attached growth on submerged objects (plants, rocks or sand and mud), preserved in 4% formaldehyde, treated for cleaning of the cell content in laboratory and preserved in permanent slides mounted in Naphrax (i.d. 1.73). Diatom species composition and abundance in the sample is determined under a light microscope (Nikon Eclipse E-800) and obtained by counting of 200 cells per sample (slide). The specific species abundance is then given as a percent of the total diatom count per sampling site (Levkov Z., 2006). Following the basic postulate (Washington, 1984) that the species composition of a given bio-indicator reflects the state of the environment in a given sample, both spatial and temporal, the dominant diatom community determined at a specific site is expected to be directly related to the measured environmental parameters (abiotic component) if a good correlation is obtained, the correlation can be used as reliable indicator.

3.2 Bio-indicators of the water quality

The data used in the study came from EU project TRABOREMA. The data covers one and a half year period, from 3.2005 to 9.2006. In total, 320 water samples were available, 224 from the lake measurement stations and 96 from the river stations, on which both physical/chemical and biological analyses were performed, the former provided the environmental variables for the habitat models, while the latter provided information on the relative abundance of the studied diatoms. The diversity indices are calculated by the mathematical formula of the indices that is used widely in the literature (Meredith, 2007).

These diatoms are more or less influenced by the following physical and chemical parameters (water properties): temperature dissolved oxygen, oxygen saturation, oxygen deficit, transparency, conductivity, pH, nitrogen compounds (NO2, NO3, NH4, Total_N, Inorganic_N, Organic_N), phosphorus compounds (Total_P), SO4, Na, K, Mg, Cu, Mn, Zn. The datasets for each diatom contain 224 measurements from the lake sampling stations.

4. DIVERSITY INDICES FOR LAKE PRESPA

The DI models presented in this paper model/predict the diatom diversity indices which are influence by the physical-chemical parameters of environment for Lake Prespa. We have selected the Shannon Diversity Index and Shannon Evenness to be the most suitable for building the models with the abiotic characteristics. These 2 diversity indices have largest correlation with the physical-chemical parameters. These four (two from WEKA two from CLUS) models will help us to find a correlation between the diatoms diversity indices and the characteristics of the sampling sites. The models with greatest correlation gain from the 10-fold cross-validation procedure will be presented below.

To build a regression trees that will reflect the diversity indices and their influence of the diatoms habitat, the M5P algorithm implemented in CLUS and WEKA are used (J.R., 1992), (M. Garofalakis, 2003), (Witten, 1999).

Table 1. Correlation coefficients gained by the M5P algorithm using CLUS and WEKA systems – Lake Single Target

Diversity Indices – Lake Data	Training set Lake	Testing set Lake
Shannon Entropy DI - CLUS	**0.7806**	**0.618**
Shannon Evenness - CLUS	**0.839**	**0.69**
Diversity Indices – Lake Data	Training set Lake	Testing set Lake
Shannon Entropy DI - WEKA	**0.6216**	**0.4942**
Shannon Evenness - WEKA	**0.645**	**0.5216**

Table 2. Correlation coefficients gained by the M5P algorithm using CLUS and WEKA systems – Rivers Single Target

Diversity Indices – River Data	Training set Rivers	Testing set Rivers
Shannon Entropy DI - CLUS	**0.8671**	**0.2141**
Shannon Evenness - CLUS	**0.8595**	**0.1891**
Diversity Indices – River Data	Training set Rivers	Testing set Rivers
Shannon Entropy DI - WEKA	**0.5502**	**0.0861**
Shannon Evenness - WEKA	**0.5412**	**0.1424**

The results from the experiments conducted on the diversity indices are given in Table 1 and 2, both from lake measurements data and rivers measurements data, respectively. Presented correlation from the both WEKA and CLUS system with 10-fold cross-validation correlation shows that CLUS system predicts the relationship between the biological and abiotic factors with greater precisions.

4.1 Diversity indices model for diatoms Shannon Evenness - WEKA system

The regression tree constructed for Shannon evenness is given on Fig 1. From the regression tree it is obvious to see that the most influence parameters on the diatoms evenness according Shannon formula is NO_3. Secondly important physical-chemical parameters are temperature and Deficit of Oxygen (DefiO). According to the generated Linear Models (LM) 1 and 6, has largest value for Shannon Evenness. They are achieved if $NO_3 > 2.718$ mg/l and DisO > -1.74 mg/l for LM1 and plus if Total Phosphorus concentration is smaller than 26,925 mg/l.

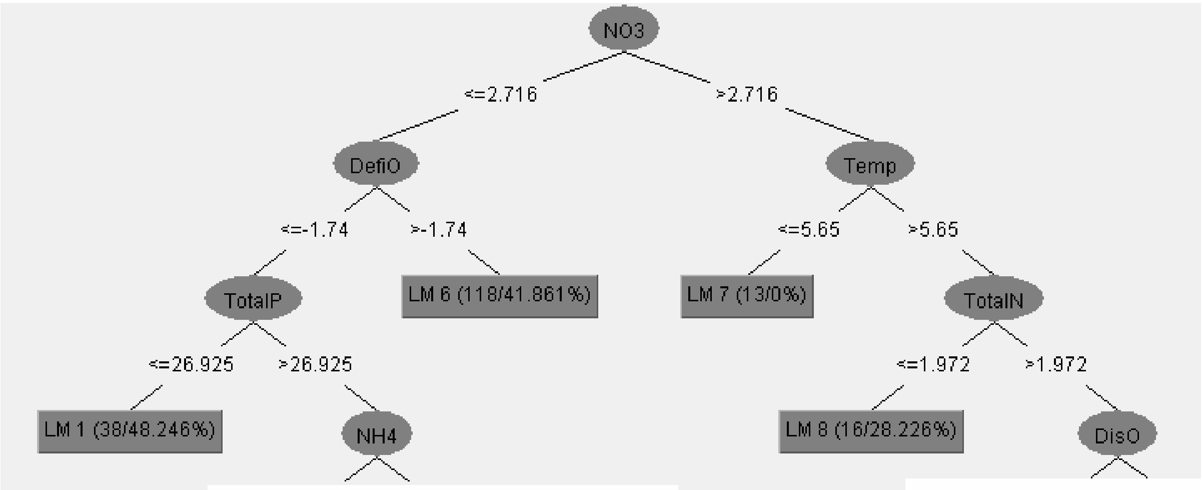

Fig. 1 – Regression tree predicting the most influence parameters on the diatoms Shannon Evenness for the Lake Prespa

The model obtained from the river measured data didn’t show build any prediction tree between this diversity index and the river physical-chemical parameters. The evaluated performances of the models are summarized in Table 2.

4.2 Diversity indices model for diatoms Shannon Evenness - WEKA system

The regression tree constructed for Shannon Evenness is given on Fig 2. From the regression tree it is obvious to see that the most influence parameters on the diatoms evenness according Shannon formula is Temperature. According to the generated regression tree model, 94% of the instances taken into account are predicted to be where temperature is higher than 5.5 °C. What this means for the index? The Shannon Diversity Index commonly used to characterize species diversity in a community, which means that if the temperature > 5.5 °C, Conductivity < 240 µS/cm and Zn concentration is lower than 17.1 mg/l and Dissolved Oxygen of the water is lower than 11.2 mg/l we have 0.75 indices of Shannon. This environment according the model is suitable from diatoms existence. In contrast, if the concentration of the Zn in greater than 17.1 mg/l we have lower score for this diversity index is 0.16. This is expected, because the Zn concentration negatively influences on the diatoms. Most of the high concentrations of the heavy metals like Zn are toxically for the diatoms. Only strict concentrations of this chemical element are allowed to coexist with the environment of the diatoms.

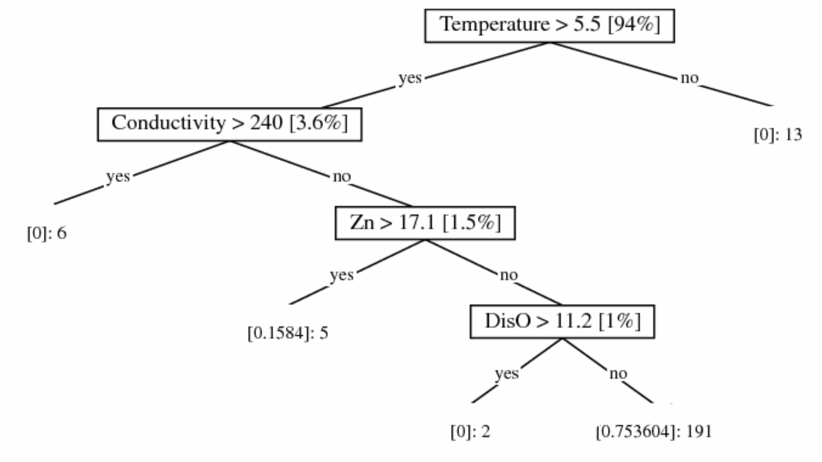

Fig. 2 – Regression tree predicting the most influence parameters on the diatoms Shannon Evenness for the Lake Prespa

Yet according to the biological expert, diatoms can be used to reconstruct patterns of water temperature, pH and values of eutrophication patterns (Svetislav, 2007).

While the models obtain from the river measured data didn't show produce any decision tree between this diversity index and the river physical-chemical parameters.

4.3 Diversity indices model for diatoms Shannon DI - WEKA system

The previous indices show the evenness of the diatoms in the lake and the rivers, while the Shannon index will present the diversity entropy of the diatoms in Lake Prespa. The generated model is shown on Fig 3.

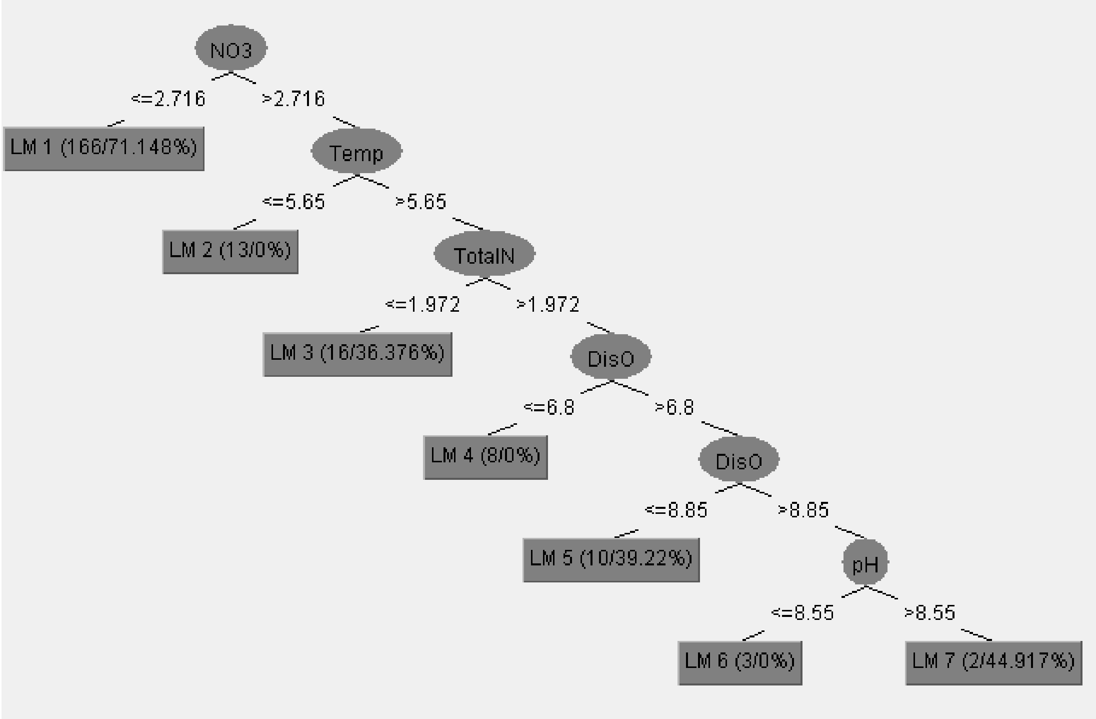

Fig. 3 – Regression tree predicting the most influence parameters on the diatoms Shannon Diversity Index for the Lake Prespa

The model shows that NO_3 is the most important environmental parameter, while the Temperature is second. This model is very similar to the case for the Shannon Evenness model. According the LM (Linear Model) prediction, LM1 have largest value, while the LM2 have the lowest value.

4.4 Diversity indices model for diatoms Shannon DI - CLUS system

The previous sections we present various indices of the Lake Prespa diatoms, while the Shannon evenness index is presented in this section. The generated model from the entire lake diatom structure is shown on Fig 4.

The model shows that temperature is the most important environmental parameter, while the Conductivity is second. This model is very similar to the previous presented indices. The lowest values for this index is 0.29 under temperature > 5.5 °C and Zn > 17.1 mg/l, while the highest values of Shannon diversity index of 2.25, if the temperature is greater than 5.5 °C. From the state above, we can conclude that positive influence on the diatom diversity indices we have if the environment temperature is greater than 5.5 °C. But, if the concentration of Zn is greater than 17.1 mg/l we have negative influence of the environment on the diatoms.

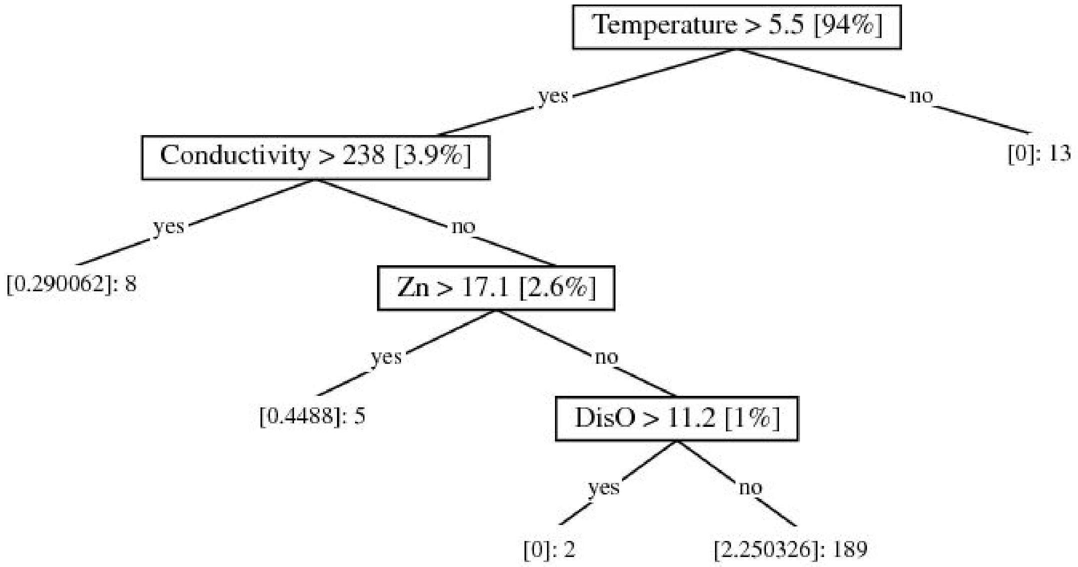

Fig. 4 – Regression tree predicting the most influence parameters on the diatoms Shannon Diversity Index for the Lake Prespa

This is expected because the Zn – heavy metal is toxic in the water for the diatoms life cycle. With this model we conclude the diversity modelling experiment and rule inductions.

5. CONCLUSION

In this paper, we applied machine learning techniques to learn a predictive model for two diversity indices of diatom species in Lake Prespa. Regression trees, which are piece-wise constant functions, are learned from measured data gained from the lake and river sampling stations, for the entire Lake Prespa, acquired within the monitoring programme of the EU project TRABOREMA. For comparison, we also developed linear regression models.

The learned models show that the most important factors influencing the diatoms diversity indices are the temperature and NO_3, while the Zn and Dissolved Oxygen (DisO) are second important. All the models given in this paper are first attempt to model the diversity indices of the diatoms in Lake Prespa. With these models we try to reveal some of the eutrophication patterns that exist in the Lake Prespa using the diatoms as bio-indocators (Moss, 1973), (Levkov, 2007).

Important to note here, that variable of the temperature and NO_3 concentration highly depends from outside factors. Nitrogen loading from the human activates - industry, while the temperature from the human activity - CO_2. As the climate models shows the temperature in the next 50 years will increase, which puts in danger existents of the diatoms according these presented models.

The experiments showed that machine learning tools can extract some valuable knowledge in a relatively comprehensible form, even when the application area is so extremely complex also for humans and the data are far from being perfect. Note that, any ecosystem cannot be fully described with all its inside process, because the model will be complex for any analysis to be performed.

Our work shows that certain rules for the diversity indices are possible to extract by using machine learning techniques. Consequently, the decision makers could improve the waste water management policy to prevent the future deterioration of the environment and the related biota.

We plan to conduct further investigation for the other diatoms in Lake Prespa and more diversity indices with more sampling data. Later we can divide the datasets via distributions: one distribution of the abundance from site to site, and other time distribution by months. In this way we will see, how the diversity indices will change in space and time. More importantly, we plan to build models predicting the structure of the community, i.e., the relative abundance for all diatom species simultaneously. For this purpose, we intend to use the machine learning methodology of multi-objective regression trees.

Bibliography

Allaby, M. (1996). *Basics of Environmental Science.* London: Routledge.

Džeroski, S. (2001). Applications of symbolic machine learning to ecological modelling. *Ecological Modelling 146* , 263-273.

Garofalakis M., D. H. (2003). Building decision trees with constraints. *Data Mining and Knowledge Discovery, 7(2)* , 187–214.

Joergensen, S. a. (2001). *Fundamentals of Ecological Modelling.* Amsterdam: Elsevier.

Levkov Z., B. S. (2007). Ecology of benthic diatoms from Lake Macro Prespa (Macedonia). *Algological Studies 124* , 71-83.

Levkov Z., K. S. (2006). Diatoms of Lakes Prespa and Ohrid (Macedonia). *Iconographia Diatomologica 16* , 603.

Meredith, M. (2007, oct 26). Retrieved November 26, 2008, from WCS Malaysia Program - Study design and data analysis: http://www.wcsmalaysia.org/stats/diversityIndexMenagerie.htm

Moss, B. (1973). The influence of environmental factors on the distribution of freshwater algae: an experimental study. IV. Growth of test species in natural lake waters and conclusions. *Journal of Ecology 61* , 193-211.

Quinlan J.R. (1992). Learning with continuous classes. *In Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence* (pp. 343-348). Singapore: World Scientific.

Reynolds. C. S. (1998). What factors influence the species composition of phytoplankton in lakes of different trophic status? *Hydrobiologia 369/370* , 11-26.

River diatoms: a multiaccess key. (n.d.). Retrieved 10 14, 2008, from http://craticula.ncl.ac.uk/EADiatomKey/html/environment.html

Svetislav, K. S. (2007). Selecting appropriate bioindicator regarding the WFD guidelines for freshwaters – a Macedonian experience. *International Journal on Algae 2007, 9(1)* , 41 – 63.

TRABOREMA Team. (2005-2007). *TRABOREMA Project WP3,*. Skopje, Macedonia: EC FP6-INCO project no. INCO-CT-2004-509177.

Van Dam H., M. A. (1994). A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. *Netherlands Journal of Aquatic Ecology 28(1)* , 117-133.

Washington H.G. (1984). Diversity, biotic and similarity indices: A review with special relevance to aquatic environment. *Water Research 18(6)* , 653-694.

Witten, I. a. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann.

Andreja Naumoski is born in 1983 in Struga, Republic of Macdonia. In 2006 he gained a University Degree from Faculty of Electrical Engineering and Information Technologies in Skopje, Macedonia as Electrical Engineering. In 2008 was awarded a Degree of MSc in Computer Science specialized in area of Eco-informatics with his MSc Title Thesis "Dynamic and habitat suitability models of Lake Prespa". He is currently working on 2 projects at the Faculty of Electrical Engineering and Information Technologies in Skopje, Macedonia. From April 2008 he start his work on his PhD Thesis titled "New classification algorithms for analysis and knowledge discovery using diatoms as bio-indicators of aquatic ecosystems" improving the ecological knowledge about the diatoms community used as bio-indicators with state-of the art information technologies, methods and algorithms of data mining methodology.