MULTI-TARGET MODELLING OF THE DIATOM DIVERSITY INDICES IN LAKE PRESPA

NAUMOSKI, A.*

University Ss. "Cyril and Methodius" - Skopje, Faculty of Electrical Engineering and Information Technologies
Karpos II bb, P.O.BOX 574, Skopje, Macedonia
(phone: + +389-2-309-9168; fax: + 389-2-306-4262)

**Corresponding author*
e-mail: andrejna@feit.ukim.edu.mk

(Received 30th June 2009; accepted 10th February 2012)

Abstract. In this paper we present models of relationship between the diatoms community diversity indices (DIs) and the physico-chemical parameters using machine learning techniques. 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 time and space. Diversity indices provide important information about rarity and commonness of species in a community. Because the physical-chemical conditions of the environmental influence on the several diversity indices of the diatoms community at once, it is more reliably to model all the diversity indices together.
For modelling of the DIs models we use the raw; as measured, values of the concentrations for the physical-chemical parameters and the diversity indices of the diatoms abundance. The well known machine learning techniques are used to express this relationship: regression trees (RTs) and multi-target regression trees (MTRT's). The MTRT are more general than the RT, which predictive target is only one variable. The diversity indices are calculated for all diatoms of one measurement for 16 months, monthly and then are placed with the given physico-chemical parameters in one table. The results from the model have captured the ecological information with correlation between 0.9 and 0.92 for unseen (test) data. Diversity indices have proved to be a reliable indicator for the influence of the environment on the diatoms community. Temperature and conductivity components together with the Zn concentration are most influenced factors on the diatoms biodiversity. This could lead to more widely research broad view in this direction of ecological modelling.
Keywords: *diatoms, Multi-target modelling, Lake Prespa, machine learning*

Introduction

Machine learning techniques into ecology have proven to be useful into obtaining knowledge for certain problems. This property is used to model the abiotic environment. Using this DIs it will be very useful to model the specific diatom communities which are known to exist only in definite environmental conditions, especially diatoms abundance which throughput last decade proved as ideal bio-indicators (Reid et al., 1995).

Taking 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. The ability to quantify diversity in this way is an important tool for biologists trying to understand community structure (Van Dam et al., 1994). Eutrophication, metals, temperature, pH, ammonia and others are one of the few that are vital for diatom survival and life cycle. This is why we build models to see how diatoms diversity indices response on the changes of these parameters.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

In order to extract this knowledge from the ecological data we will 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 (Levkov et al., 2006). 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.

From machine learning methodology we explore the two afore mentioned possibilities for habitat modelling of the diatom community in Lake Prespa (Republic of Macedonia). To learn a model for each diatom species separately we employ regression trees – STRT (Breiman et al., 1984). To build a model for the all the diversity indices, we use multi-target regression trees – MTRT (Blockeel et al., 1998; Struyf et al., 2006). The main advantages of the latter approach are: (1) the multi-target model is smaller and faster to learn than learning models for each organism separately and (2) the dependencies between the organisms are explicated and explained.

We use dataset, which has been collected from different measurement stations placed in Lake Prespa, as a part of the EU project TRABOREMA (WP3, EC FP6-INCO project no, 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.

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 model for Lake Prespa, in Section 2 we give an overview of the machine learning algorithms and introduction to machine learning for diversity indices modelling, with a briefly description of the approach to machine learning that is often used in this kind of modelling: multi-target regression trees. The measured data and experimental setup, 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 indices and single-target only. Section 5 concludes the paper and gives directions for future work in this area.

Machine learning algorithms

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 calculated with mathematical equations which are known the literature (http://www.wcsmalaysia.org/stats/diversityIndexMenagerie.htm).

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

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.

Diversity 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 [11]. We are interested to model the diatoms community, which latter can be used as bio indicator of the environment. In these experiments we have calculated Chao richness, Hill's N1, Hill's N2, $(\text{Berger-Parker})^{-1}$, $(\text{Simpson})^{-1}$, ShannonH' - Entropy, Brillouin, Margalef, Hill's N2/N1, Brillouin Evenness, Simpson Evenness and Shannon Evenness [11].

Nevertheless, we have taken into the experiment several other indices, the results are provided later in this paper. The influences of the abiotic factors on the community structure are captured using the both single and multi regression trees.

Multi-target regression trees

Multi-objective regression trees generalize regression trees in the sense that they can predict a value of multiple numeric target attributes (Breiman et al., 1984). Therefore, as prediction, instead of storing a single numeric value, the leafs of a multi-objective regression tree store a vector. Each component of this vector is a prediction for one of the target attributes.

A multi-objective regression tree (and a regression tree (Blockeel, H., et. al., 1998)) is usually constructed with a recursive partitioning algorithm from a training set of records (known as algorithm for top-down induction of decision trees). The records include measured values of the descriptive and the target attributes. One of the most important steps during the tree induction algorithm is the test selection procedure. Each test for a given node is selected on the base of some heuristic function that is computed

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

on the training data (Garofalakis et al., 2003). The goal of the heuristic is to guide the algorithm towards small trees with good predictive performance.

After the regression tree is constructed, it is common to prune it. With pruning some subtrees are replaced with leaves, in order to improve predictive accuracy and/or interpretability. There are two pruning approaches: pre-pruning and post-pruning. With pre-pruning approaches, the pruning is included in the tree building algorithm as a stopping criterion. Examples of pre-pruning are the stopping criteria mentioned above: the number of records in a leaf and the maximum depth of the tree. The post-pruning approaches are applied after the tree construction has ended. Example of this approach is the pruning method proposed by (Blockeel et al., 2002). Essentially, this is a dynamic programming optimization method that selects a subtree from the constructed tree with at most *maxsize* nodes and minimum training set error (mean squared error, summed over all target attributes). The restriction *maxsize* is a user defined value.

Data description

The data that we have at hand were measured during the EU project TRABOREMA. The measurements cover one and a half year period (from March 2005 till September 2006). Samples for analysis were taken from the surface water of the lake at several locations near the mout of the major tributaries. In total, 275 water samples were available, 218 from the lake measurements and 57 from the tributaries. On these water samples both physico-chemical and biological analyses were performed. The physico-chemical properties of the samples provided the environmental variables for the habitat models, while the biological samples provided information on the relative abundance of the studied diatoms.

The following physico-chemical properties of the water samples were measured: temperature, dissolved oxygen, Secchi depth, conductivity, pH, nitrogen compounds (NO2, NO3, NH4, inorganic nitrogen), SO4, and Sodium (Na), Potassium (K), Magnesium (Mg), Copper (Cu), Manganese (Mn) and Zinc (Zn) content.

The biological variables were actually the relative abundances of 116 different diatom species. Diatom cells were collected with a planktonic net or as an attached growth on submerged objects (plants, rocks or sand and mud). This is the usual approach in studies for environmental monitoring and screening of the diatom abundance [8]. The sample, afterwards, is preserved and the cell content is cleaned. The sample is examined with a microscope, and the diatom species and abundance in the sample is obtained by counting of 200 cells per sample. The specific species abundance is then given as a percent of the total diatom count per sampling site (WP3, EC FP6-INCO project no, 2005-2007). The diversity indices are calculated by the mathematical formula of the indices that are used.

We applied the methodology described in Section 2, according to the experimental setup describe in next section, to the data at hand. With the modelling procedure (with the different scenarios and the different pruning algorithms) we obtained several models. From these models we select the ones that have better predictive power, and have reasonable size (in the most cases the tree minimal records in leaf is 4). The diatom species in the models are presented with their respective abbreviations. Their complete names can be found in (Levkov et al., 2006).

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

With the multi-target modelling we obtain predictive model that describe all diatom species and explains the dependencies between them and the physical-chemical characteristics.

Experimental design

We perform the analysis along two different scenarios: (1) modelling with the raw, as measured data and (2) selected bio-diversity indices from the previous experiment which have best correlation coefficient and RMSE. These scenarios were applied on lake measurements data for all 18 physico-chemical parameters and 12 biodiversity indices. In the case when as target variables we had all diatoms diversity indices we learn single MORT and a regression tree or single-target regression tree (STRT) for each index separately.

We applied 3 different pruning algorithms: minimal records in a leaf, maximal depth and maximal size. The parameter setting for these algorithms was as follows: for minimal records in a leaf we set 2, 4, 8, 16 and 32; for maximal depth we set 3, 4 and 5 and for maximal size we set 7, 9, 11 and 13. For validation of the performance on unseen data we used 10-fold cross validation. To assess the predictive power of the models we compare them by their correlation coefficient and root mean squared error (RMSE).

Diversity indices models of diatoms community

Diversity indices models

The MTRT tree constructed from the raw data is given on *Fig 1.* From the tree it is obvious to see that the most influence parameters on the diatoms diversity indices is the temperature and specifically the temperature high than 5.5 °C. Secondly important physico-chemical parameters are conductivity and the Zn. If we inspect the model leafs of the tree we can note that largest values are encounter if the temperature is high than 5.5 °C and conductivity values are low than 238 μS/cm.

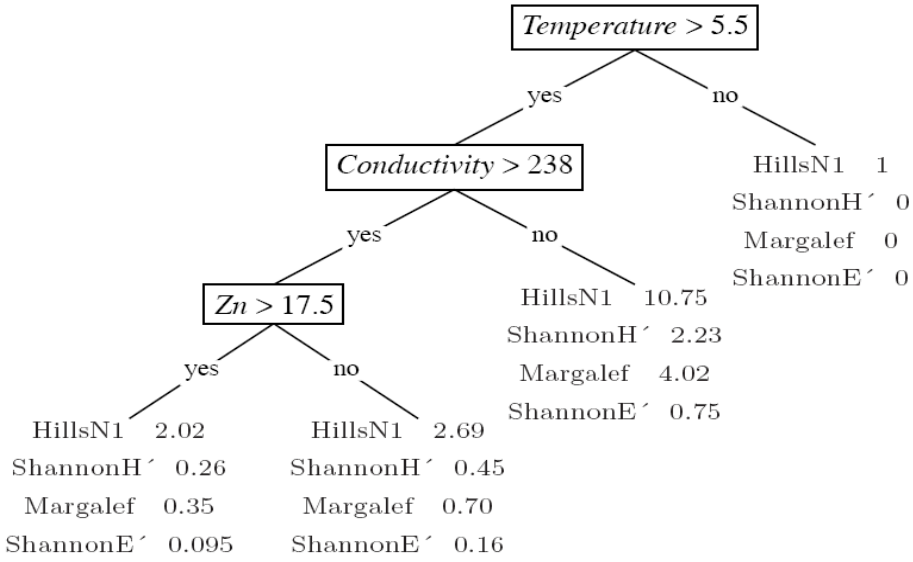

Figure 1. *MORT of the TOP 10 Diatom from lake measurements dataset*

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

This is expected; in fact the diatoms are known species that leaves in moderate water with low level of toxic element as the model shows. The Zn concentration negatively influences on the diversity of the diatoms, also the high temperatures. The constructed model for the measured data consist from the 4 important diversity indices, which according to the model are the most influence factors on the diatoms community. It will be very interesting to see how these indices are affected separately by the 18 physical-chemical parameters. For this purpose we build single-target regression trees – STRT.

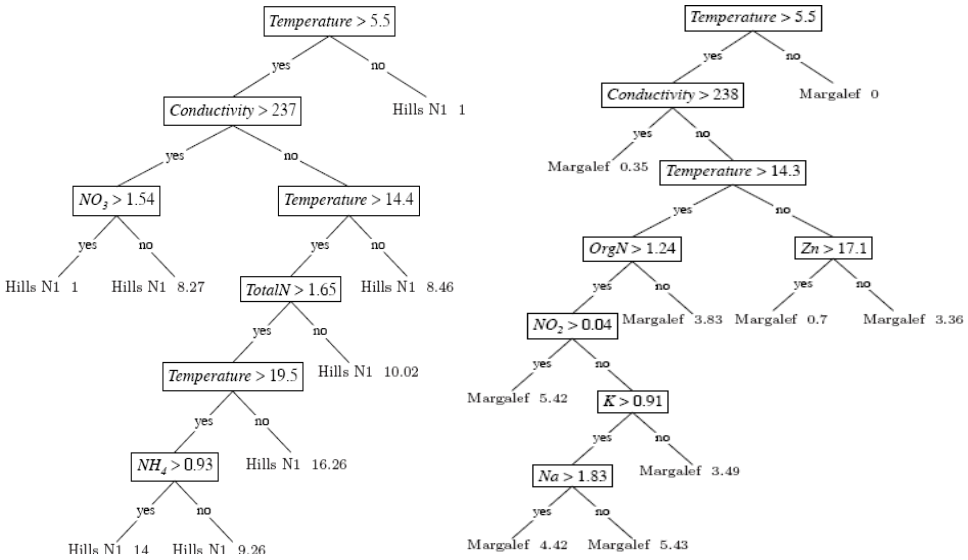

Figure 2. *Single target regression tree for the Hills N1 (left) and Margelef (right) diversity indices*

On *Fig. 2* - left, the single target regression tree for the Hills N1 diversity index is shown. It is obvious that the temperature is the most influence factor on this index, as the MTRT have shown. Beside the temperature and the conductivity, other influence factors are nitrogen components. Total Nitrogen, NO3 and NH4 play important role in the organisms' life cycle. High values of this index are expected in relative high temperatures, low conductivity and low concentrations of nitrogen components. Later in the experiment we build a model for single target model for Margelaf diversity index (see *Fig. 2* – right side). It is obvious again that the temperature is the most influence factor on this index and conductivity as second. Beside the temperature and the conductivity, other influence factors are nitrogen components. Total Nitrogen, NO3 and NH4 together with some metal parameters. Here it is interesting that metals like Na, K and Zn have influence on this index, because the Zn especially have toxic property on the environment.

High values of this index are expected in relative high temperatures, low conductivity. The nitrogen component together with the metals it is obvious that do not change much the value of this index.

Single target model was also build for the Shannon Evenness and Shannon H' - Entropy index. The single target tree for the Shannon H' tree is shown on *Fig. 3* – left side. As the rest of the model this index is mostly influence by the temperature and the conductivity parameter together with a set of nitrogen components. The Zn component capture with the multi-target regression tree has also has showed here. Low values of

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

this index are expected with high values of the temperature and conductivity, which was confirmed by the MTRT for all indices.

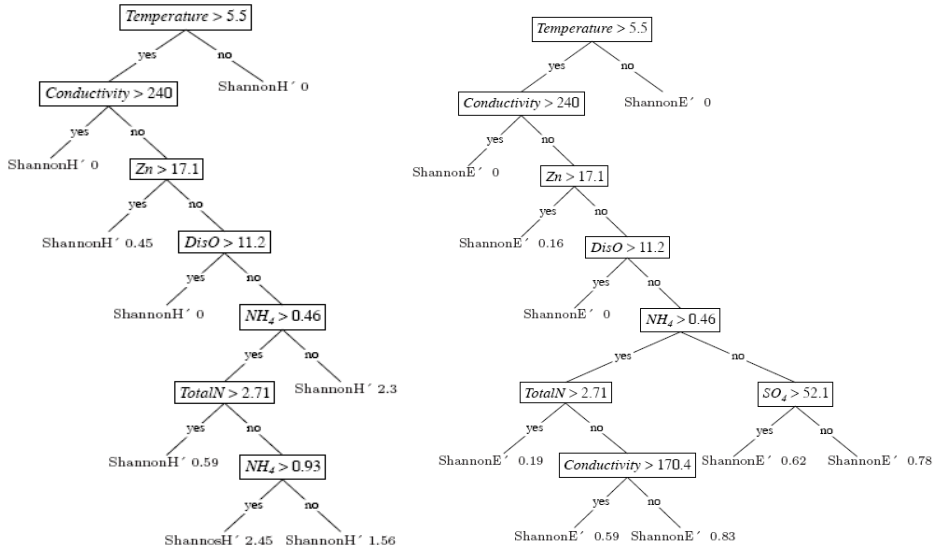

Figure 3. *Single target regression tree (STRT) for the Margelaf and Shannon H' diversity index*

Evenness property of the diatoms community is presented with Shannon E' index, which we can see from the picture the most influence parameter, is the temperature and the conductivity (*Fig. 3* – right side).

Table 1. *Correlation coefficient and RMSE coefficients for the diversity indices for both STRT and MTRT*

	MTRT					**STRT**			
	CC		**RMSE**			**CC**		**RMSE**	
	Train	**Xval**	**Train**	**Xval**		**Train**	**Xval**	**Train**	**Xval**
Chao richness	0.00	0.00	7.84	13.96		0.00	0.00	3.29	13.87
Hill's N1	0.94	0.22	1.92	6.88		0.98	0.27	1.06	6.88
Hill's N2	0.00	0.00	1.38	5.56		0.00	0.00	1.03	6.21
1 /Berger-Parker	0.00	0.00	0.66	2.55		0.00	0.00	0.66	2.56
1 / Simpson	0.00	0.00	1.77	7.16		0.00	0.00	1.74	7.65
ShannonH'	0.90	0.42	0.39	0.94		0.98	0.47	0.20	0.94
Brillouin	0.00	0.00	0.18	0.63		0.00	0.00	0.16	0.74
Margalef	0.89	0.40	0.81	1.84		0.98	0.46	0.38	1.82
Hill's N2/N1	0.00	0.00	0.05	0.17		0.00	0.00	0.03	0.19
Brillouin Evenness	0.00	0.00	0.06	0.20		0.00	0.00	0.05	0.21
Simpson Evenness	0.00	0.00	0.07	0.25		0.00	0.00	0.08	0.26
Shannon Evenness	0.87	0.46	0.14	0.28		0.98	0.63	0.06	0.25

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

From the model, other influence factors are the nitrogen components and the Zn parameters, and now together with the SO_4 component plays an important role in change of this index. High values of this index are expected for low temperatures, conductivity and total nitrogen component. The SO_4 component has little vary this index.

In continue of this paper, the performances of the experiments are given in *Table 1* and *Table 2*. The correlation coefficients given in table 1 are for both MTRT and the STRTs. Many of the diversity indices do not have any correlation, which later we will dismiss them from further investigation, leading to the second experimental setup. The performance data of the experiments conducted on the 4 left diversity indices in the second experiments are given in *Table 2*.

It is obvious from the table that the experiments on the unseen data have correlation coefficient 0.6 for multi-target trees and 0.63 for the single target trees. But this do not mean that the multi-target tree approach is better, but overall power prediction of the multi-target tree are better with mean value of the four parameters of 0.55 while STRT have only 0.41. Also the multi-target tree has produced a decision tree which is smaller in size and takes all the diversity indices at once.

Table 2. *Correlation coefficient and RMSE coefficients for the diversity indices for both STRT and MTRT only for best four parameters*

	MTRT					**STRT**			
	CC		**RMSE**			**CC**		**RMSE**	
	Train	**Xval**	**Train**	**Xval**		**Train**	**Xval**	**Train**	**Xval**
Hill's N1	0.91	0.44	2.37	5.87		0.98	0.27	1.06	6.88
ShannonH'	0.92	0.58	0.36	0.82		0.98	0.47	0.20	0.94
Margalef	0.91	0.57	0.74	1.60		0.98	0.46	0.38	1.82
Shannon Evenness	0.90	0.60	0.12	0.25		0.98	0.63	0.06	0.25

Conclusion

In this paper, we applied machine learning methodology, in particular single regression trees and multi-target regression trees, to model the abiotic influence of the environment on the diversity indices of the diatom community abundance in Lake Prespa. The models of the lake diatom communities have different structure and different environmental preferences and they interact with those different diversity indices according the models.

The learned diversity indices models show that the most important factors influencing the diatoms diversity indices are the temperature and conductivity, while the Zn and the nitrogen components are second important (from the Single-target RT). All the models given in this paper are first attempt to model the diversity indices of the diatoms in Lake Prespa.

Important to note here, that variable of the temperature and conductivity concentration, together with the Zn component highly depends from outside factors. Zn concentration loading from the human activates – industry, while the temperature from the human activity – CO_2. According the climate models the temperature in the next 50 years will increase, the models indicate and puts in danger existents of the diatoms that depends from the temperature factor.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)

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. The predictive power (testing procedure) of the models is weak, but the knowledge representation (training procedure) has reach values of 0.9 in some cases. We have in mind that any ecosystem cannot be fully described with all its inside process, because the model will be complex for any analysis to be performed.

We do strongly believe that these models will help explaining the very complex environmental patterns of influence within the Prespa Lake ecosystem and emphasize the most important variables to be monitored or put in the focus of the decision makers regarding the mitigation of the detected forced eutrophication processes and their consequences. In this direction, an expert system that will automatically generate decisions rules will point out the relationship between the eutrophication parameters (like Secchi Disk, Total Phosphorus, Total Nitrogen and etc.) and the diatom community. For further work, we plan to model the diatom communities using multi-label classification methods.

Acknowledgements. This research project was funded by the bilateral project between Slovenia and Macedonia: Knowledge Discovery for Ecological Modelling of Lake Ecosystems.). The author thank for the support.

REFERENCES

[1] Blockeel, H., De Raedt, L., Ramon, J. (1998): Top-down induction of clustering trees. – In: Shavlik, J. (Ed.), Proceedings of the 15th International Conference on Machine Learning, pp. 55-63.

[2] Blockeel, H., Struyf, J. (2002): Efficient algorithms for decision tree cross-validation. – Journal of Machine Learning Research 3: 621-650.

[3] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984): Classification and Regression Trees. – Wadsworth.

[4] Garofalakis, M., Hyun, D., Rastogi, R., Shim, K. (2003): Building decision trees with constraints. – Data Mining and Knowledge Discovery 7(2): 187-214.

[5] Levkov, Z., Krstic, S., Metzeltin, D, Nakov, T., (2006): Diatoms of Lakes Prespa and Ohrid (Macedonia). – Iconographia Diatomologica 16: 603 pp.

[6] Reid, M.A., Tibby, J.C., Penny, D., Gell, P.A. (1995): The use of diatoms to assess past and present water quality. – Australian Journal of Ecology 20(1): 57-64.

[7] Struyf, J., Dzeroski, S. (2006): Constraint based induction of multi-objective regression trees, Knowledge Discovery in Inductive Databases. - 4th International Workshop, KDID'05, LNCS vol. 3933: 222-233.

[8] TRABOREMA Project WP3, EC FP6-INCO project no. INCO-CT-2004-509177 (2005-2007)

[9] Van Dam, H., Martens, A., Sinkeldam, J. (1994) A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. Netherlands Journal of Aquatic Ecology 28(1): 117-133.

[10] WFD Water Quality - Sampling – Part 2: (1993): Guidance on sampling techniques. – (ISO 5667-2:1991)

[11] http://www.wcsmalaysia.org/stats/diversityIndexMenagerie.htm, Access 26 November 2008.

http://www.ecology.uni-corvinus.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)