PREDICTING CHEMICAL PARAMETERES OF THE WATER FROM DIATOM ABUNDANCE IN LAKE PRESPA AND ITS TRIBUTARIES

Andreja Naumoski[1], Dragi Kocev[2], Nataša Atanasova[3], Kosta Mitreski[1], Svetislav Krstić[4], Sašo Džeroski[2]

[1]Faculty of Electrical Engineering and Information Technology, Skopje, Macedonia; Andreja.Naumoski@feit.ukim.edu.mk, komit@feit.ukim.edu.mk
[2]Dept. of Knowledge Technologies, Jožef Stefan Institute Jamova 39, 1000 Ljubljana, Slovenia, Dragi.Kocev@ijs.si, Saso.Dzeroski@ijs.si
[3] Institute of Sanitary Engineering, Faculty of Civil Engineering, University of Ljubljana, Hajdrihova 28, 1000 Ljubljana, Slovenia, natanaso@fgg.uni-il.si
[4]Institute of Biology, Faculty of Natural Sciences and Mathematics, Skopje, Macedonia, skrstic@pmf.ukim.mk

Abstract
In this work, we are modelling the physic-chemical parameters of water using bioindicator data (diatom taxa abundance data). Chemical status of the water (or water quality class) is defined by the values of measured physic-chemical parameters. Traditional approach to model these data is to learn a separate model for each parameter and then derive a global overview with some kind of summarization over the multiple models. Another approach is to learn a single model that describes all parameters (multi target approach). We explore these approaches and apply them on data from Lake Prespa and its tributary rivers. The obtained models revealed interesting connections between the diatom taxa and the water quality (i.e. the values of the chemical parameters).
Key-words: Lake Prespa, diatoms, physic-chemical status, machine learning

1 Introduction

High population densities and the multiplicity of industrial and agricultural activities expose most hydrographical basins close to urban centres to heavy and rising environmental impacts. Usual approaches to water quality evaluation are divided in two main categories. One based on physical and chemical methods, and another considering biological community's evaluation [1]. Physical and chemical monitoring reflects only instantaneous measurements, restraining the knowledge of water conditions to the moment when the measurements were performed. Biotic parameters on the other hand provide better evaluation of environmental changes, because community development integrates a period of time reflecting conditions that might not be anymore present at the time of sampling and analysis.

The water quality models (WQMs) are focused on predicting water chemical parameters using the biological characteristics of the water as indicators of ecological status of the lake. Diatoms are usually considered as very reliable bioindicators of the environment [10, 11]. The relationship between the presence/abundance of these diatoms and the specific abiotic factors can be studied using machine learning techniques. This is done under the implicit assumption that both are observed at a single point in time for a given spatial unit. WQMs takes into account only the specified target abiotic factors of the environment, but still some temporal aspects may be taken into account.

In this work, we model the physic-chemical (abiotic) parameters of the water in Lake Prespa and its tributaries. We are modelling these parameters using the presence/abundance of some diatom taxa (biological parameters). The obtained models for the chemical parameters can be further used to define the water classes. We decided to use trees as modelling technique. This decision was made because the interpretability of the trees and their relatively good performance.

Having in mind that there are multiple chemical parameters (in statistical terminology - multiple response variables), we investigate two approaches for modelling: (1) learn a regression tree (RT) [2] for each chemical parameter separately and (2) learn a multiple targets regression tree (MTRT) [3, 4] to predict all parameters simultaneously. The advantages of the latter approach are that the obtained MTRT is smaller than the sum of the RTs for each chemical parameter and that the MTRT is more reliable in to reveal and explain the dependencies between the different physic-chemical parameters [4].

The data that we use were collected during the EU funded project TRABOREMA (FP6-INCO-CT-2004-509177). They describe the diatom abundance in Lake Prespa and its tributary rivers [13]. The measurements comprise several important parameters that reflect the physical, chemical and biological aspects of the water quality of the lake [5, 6]. These include measurements of the relative abundance of different taxa belonging to the group Bacillariophyta (diatoms). The focus of this paper is the investigation of the relationship between their relative abundance and the abiotic characteristics of the habitat. Later, these diatoms are used as bioindicators, primary attributes for building the WQMs.

The remainder of this paper is organized as follows. In Section 2, we describe the machine learning methodology that was used (regression trees and multiple targets regression trees). Section 3 describes the data and Section 4 explains the experimental design that was employed to analyse the data at hand. In Section 5, we present the obtained WQ models and discuss them. Section 6 gives the main conclusions.

2. Methodology

2.1 Regression Trees

Regression trees are decision trees that are capable of predicting the value of a numeric target variable [2]. Regression trees are hierarchical structures, where the internal nodes contain tests on the input attributes. Each branch of an internal test corresponds to an outcome of the test, and the prediction for the value of the target attribute is stored in a leaf. Regression tree leafs contain constant values as predictions for the target variable (they represent piece-wise constant functions). To obtain the prediction for a new data record, the record is sorted down the tree, starting from the root (the top-most node of the tree). For each internal node that is encountered on the path, the test that is stored in the node is applied, and depending on the outcome of the test, the path continues along the corresponding branch (to the corresponding subtree). The resulting prediction of the tree is taken from the leaf at the end of the path. Example of regression tree is shown in Figure 4.

2.2 Multi Targets Regression Trees

Multiple Targets Regression Trees (MTRTs) generalize regression trees in the sense that they can predict a value of multiple numeric target attributes [3]. Therefore, for prediction, instead of storing a single numeric value, the leafs of a MTRT store a vector. Each component of this vector is a prediction for one of the target attributes.

A MTRT (and a RT) 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 on the training data. The goal of the heuristic is to guide the algorithm towards small trees with good predictive performance.

In this paper, we apply the system CLUS [7] for constructing the (Multiple Targets) regression trees. CLUS uses the sum of the variations in the induced subsets (intra-subset variance) as heuristic for selection of the tests. The intra-subset variance is measured as:

$$\sum_{j=1}^{T} \sum_{i=1}^{N} (x_{ij} - x'_j)^2 \tag{1}$$

where T is the number of target attributes, N is the number of records in the subset, x_{ij} the value of target attribute j of the i-th record in the subset, and x'_j the subset mean of attribute j. Lower intra-subset variation results in predictions that are more accurate.

After the regression tree is constructed, it is common to prune it. With pruning some subtrees are replaced with leafs, 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 [8]. 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.

3. 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) [5]. Samples for analysis were taken from the surface water of the lake at several locations near the mouth of the major tributaries (see Figure 1). In total, 275 water samples were available, 218 from the lake measurements and 57 from the tributaries. The three tributaries during the funded project were analysed in a lesser extent than the lake itself, but they are very important factor of ecological influence for the diatom communities. On these water samples both physic-chemical and biological analyses were performed.

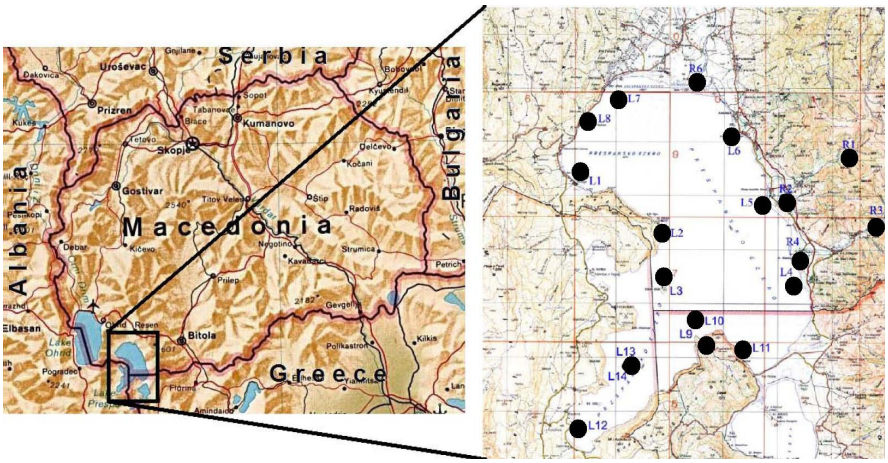

Fig. 1. Sampling locations

The physic-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 physic-chemical properties of the water samples were measured: temperature, dissolved oxygen, Secchi depth, conductivity, pH, nitrogen compounds (NO_2, NO_3, NH_4, inorganic nitrogen), SO_4, and Sodium (Na), Potassium (K), Magnesium (Mg), Copper (Cu), Manganese (Mn) and Zinc (Zn) content. Some basic statistic about the chemical parameters is presented in Table 2.

The biological variables were actually the relative abundances of 116 different diatom taxa. 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 [9]. 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 [5, 6].

From the 116 diatom taxa we selected the top 10 most abundant ones (separately for the lake and river measurements). These taxa were used to construct models for the chemical parameters.

Table 1. The 10 most abundant diatoms for lake and tributary rivers samples and their acronyms

	Lake		Rivers
Acronym	Diatom	Acronym	Diatom
APED	Amphora pediculus	AMSS	Achnanthidium minutissimum
CJUR	Cyclotella juriljii	CPLA	Cocconeis placentula
COCE	Cyclotella ocellata	DMES	Diatoma mesodon
CPLA	Cocconeis placentula	EMIN	Encyonema minutum
CSCU	Cavinula scutelloides	ESLS	Encyonema silesiacum
DMAU	Diploneis mauleri	FCAP	Fragilaria capuccina
NPRE	Navicula prespanense	HARC	Hannea arcus
NROT	Navicula rotunda	NGRG	Navicula gregaria
NSROT	Navicula subrotundata	NLAN	Navicula lanceolata
STPNN	Staurosirella pinnata	PLLA	Planothidium lanceolatum

4. Experimental Design

We model the data from the measurements at lake and tributaries sampling points. We constructed a MTRT to predict all chemical parameters simultaneously (single model for all, see Figures 2 and 4) and RT for each chemical parameter separately. In this paper we show and discuss the MTRTs and single RTs both for lake and river sites (Figure 3 and Figure 5).

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.

We present the correlation coefficients obtained on the training set, because the main purpose of the tree is to describe the relations between the measured data and not to be used for predictions.

5. Models of chemical parameters for water quality

We applied the methodology described in Section 2, according to the experimental setup, to the data at hand. With the modelling procedure (with the different scenarios and the different pruning algorithms) we obtained several models.

Table 2. Simple statistics for the chemical parameters for both datasets

	Lake				Rivers			
	Minimum	Maximum	Mean Value	Standard Deviation	Minimum	Maximum	Mean Value	Standard Deviation
Temperature [°C]	2.9	26.8	15.56	6.61	6.2	21	12.29	4.75
Dissolved Oxygen [mg/dm^3]	0.7	12.6	8.04	1.99	3.6	11	7.80	1.81
Saturated Oxygen [%]	6.6	114.19	83.07	19.54	39.4	102.5	74.40	13.15
Deficit Oxygen [mg/dm^3]	-9.32	1.33	-1.73	2.02	-5.53	0.24	-2.65	1.27
Secchi Depth [m]	1.8	5.4	3.09	0.76	/	/	/	/
Conductivity [µS/cm]	142.5	318	196.23	27.84	23	224	58.21	41.81
pH factor	5.5	24.8	8.68	2.86	5.48	8.8	6.99	0.62
NO_2 [mg/dm^3]	0	0.44	0.03	0.05	0	0.34	0.04	0.06
NO_3 [mg/dm^3]	0	13.4	2.07	2.13	0	20.71	3.25	3.55
NH_4 [mg/dm^3]	0.01	1.07	0.29	0.18	0.03	0.78	0.34	0.19
Total N [mg/dm^3]	0.09	9.07	2.07	1.12	0.24	4.59	1.85	1.17
Inorganic N [mg/dm^3]	0.01	0.83	0.22	0.14	0.02	0.61	0.26	0.15
Organic N [mg/dm^3]	0.02	8.41	1.83	1.10	0.08	4.32	1.59	1.10
SO_4 [mg/dm^3]	2.68	266.1	29.47	22.98	1.18	102.3	25.70	26.75
Total P [µg/dm^3]	1.15	83.13	18.63	15.31	1.5	125	18.58	20.96
Na [mg/dm^3]	0.75	13.15	4.36	2.10	0.73	8.89	2.09	1.32
K [mg/dm^3]	0.23	4.8	1.50	0.66	0.31	6.65	1.19	1.04
Mg [µg/dm^3]	1.11	19.45	5.70	2.84	0.23	9.63	2.50	2.50
Cu [µg/dm^3]	1.04	23.3	3.97	2.79	0.64	13.28	4.43	3.00
Mn [µg/dm^3]	0.88	230	7.88	16.79	1.04	79.3	16.51	19.25
Zn [µg/dm^3]	0.27	22.7	5.23	4.42	0.25	214.5	9.84	29.48

From these models we select the ones that have better predictive power, and have reasonable size (in the most cases the tree size is 9). The diatom species in the models are presented with their respective abbreviations. Their complete names can be found in Table 1.

Table 3. Correlation Coefficient obtained on the training sets. **MTRT** = Multi Target Regression Trees, **RT** = Regression Trees

	Lake		Rivers	
	MTRT	RT	MTRT	RT
Temperature	0.5	0.58	0.65	0.80
Dis. Oxygen	0.52	0.33	0.63	0.84
Sat. Oxygen	0.53	0.54	0.47	0.78
Def. Oxygen	0.56	0.57	0.41	0.78
Secchi Depth	0.33	0.50	/	/
Conductivity	0.32	0.55	0.73	0.86
pH Factor	0.19	0.77	0.46	0.73
NO_2	0.44	0.62	0.64	0.81
NO_3	0.51	0.64	0.51	0.64
NH_4	0.26	0.45	0.79	0.63
Total N	0.35	0.44	0.65	0.78
Inorganic N	0.26	0.43	0.79	0.86
Organic N	0.32	0.44	0.63	0.75
SO_4	0.2	0.67	0.55	0.87
Total P	0.25	0.53	0.73	0.80
Na	0.27	0.43	0.66	0.84
K	0.32	0.41	0.71	0.76
Mg	0.36	0.43	0.55	0.82
Cu	0.21	0.46	0.33	0.80
Mn	0.1	0.31	0.65	0.92
Zn	0.33	0.42	0.30	0.71

The correlation coefficients of the obtained trees are presented in Table 3 and 4. We can note that the performance of the trees from river measurements is generally better than the performance of the trees from lake measurements. Also, the regression trees achieve slightly better performance than the multi target regression trees.

Table 4. Correlation Coefficient obtained on the testing sets. **MTRT** = Multi Target Regression Trees, **RT** = Regression Trees

	Lake		Rivers	
	MTRT	RT	MTRT	RT
Temperature	0.25	0.38	0.1	0.4
Dis. Oxygen	0.14	0.03	0.09	0.44
Sat. Oxygen	0.06	0.41	0.11	0.17
Def. Oxygen	0.02	0.26	0.23	0.16
Secchi Depth	0.04	0.06	/	/
Conductivity	0.14	0.35	0.02	0.09
pH Factor	0.04	0.02	0.21	0.02
NO_2	0.3	0.2	0.05	0.12
NO_3	0.35	0.35	0.05	0.21
NH_4	0.03	0.06	0.36	0.3
Total N	0.09	0.2	0.06	0.14
Inorganic N	0.03	0.09	0.36	0.3
Organic N	0.07	0.17	0.1	0.05
SO_4	0.03	0.11	0.29	0.34
Total P	0.03	0.13	0.09	0.15
Na	0.02	0.16	0.02	0.31
K	0.01	0.06	0.15	0.22
Mg	0.14	0.12	0.1	0.24
Cu	0.02	0.03	0.35	0.02
Mn	0.04	0.1	0.22	0.32
Zn	0.1	0.05	0.1	0.02

5.1 Models from lake measurements

Figure 2 depicts a MTRT that presents the relationship between the presence/absence of diatom taxa and the abiotic parameters. This tree illustrates 6 different chemical situations that are described in the leafs of tree and focus the analysis on the trophic state indicators (total phosphorus, secchi depth and nitrogen). The situation when *Navicula rotunda* (NROT) and *Diploneis mauleri* (DMAU) occurs with situation when NROT is present and DMAU in smaller quantities (the two left-most leafs from the tree). We can conclude that DMAU indicates chemical situation with lower total phosphorus and nitrogen (indicators of clean water status) and, correspondingly, higher Secchi depth values. Absence of NROT and *Cyclotella ocellata* (COCE) (right most leaf) indicates higher phosphorus values (between eutrophic and mesotrophic state).

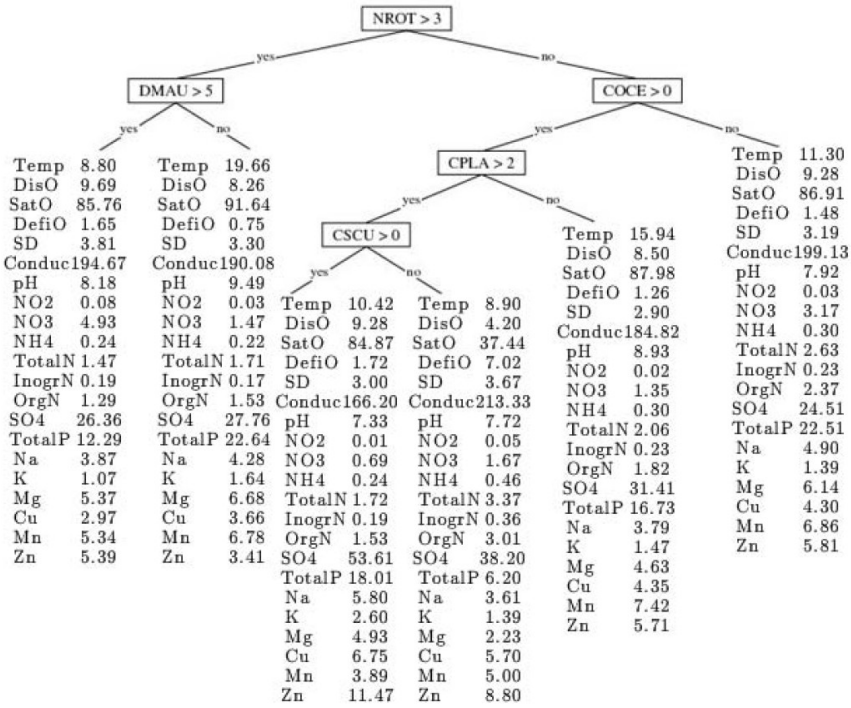

Fig. 2. MORT of the TOP 10 Diatom from lake measurements dataset

Lowest phosphorus values can be related to absence of NROT and *Cavinula scutelloides* (CSCU) and presence of COCE and *Cocconeis placentula* (CPLA) (oligotrophic state). The rest of the situations indicate similar trophic state. Additionally, this tree shows that NROT is capable of living in alkaline environment (higher pH levels).

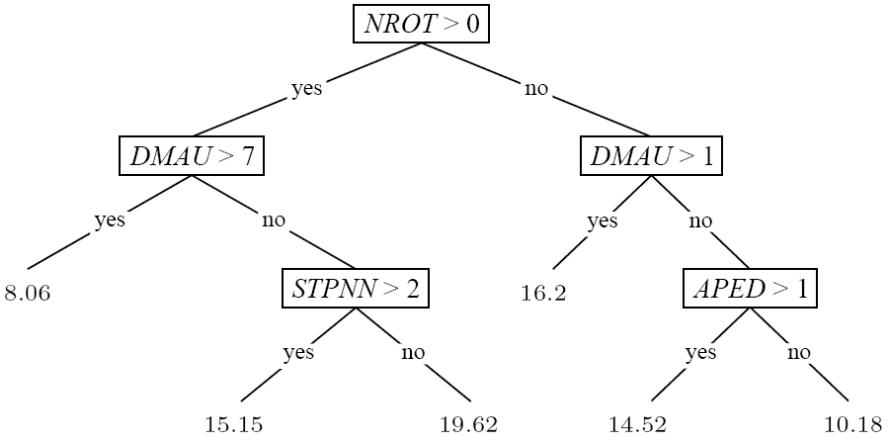

Fig. 3. Regression tree for temperature from lake measurements

The tree presented in Figure 3 shows similar relations between the temperature and the diatoms as its more general variant from Figure 2. It is easy to see that an NROT diatom is mostly influenced by the temperature. High temperatures are favourable environment for NROT, but low temperature is satisfying for DMAU population if higher than 7.

5.2 Models from river measurements

Figure 3 presents a tree obtained from the data from the tributary rivers. It represents chemical situations that are generalized from all tributary rivers.

The presence of *Navicula lanceolatum* (NLAN) (left most leaf) indicates the highest total phosphorus concentrations and higher metal concentrations as compared to the other chemical situations from the tree. If there are relatively low quantities of NLAN (6 - 21) and higher quantities of *Navicula gregaria* (NGRG) then the conductivity is high and the concentration of nitrates (NO_2 and NO_3) is higher.

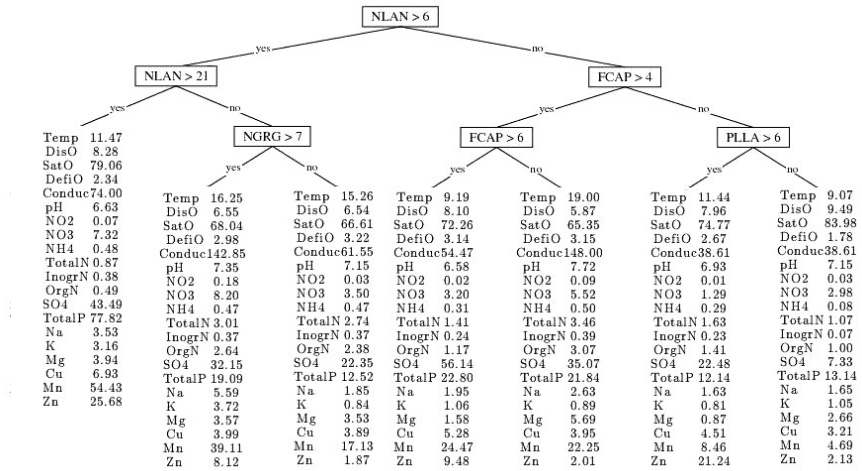

Fig. 4. MORT of the TOP 10 Diatom from inflow rivers measurements dataset

Highest conductivity can be found at lower or no presence of NLAN (less than 6) and presence of *Fragilaria capucina* (FCAP) (between 4 and 6). Higher concentration of FCAP, in this case is encountered at lower temperatures. Relatively low concentrations of phosphorus can be found at low or no presence of NLAN (less than 6) and FCAP (less than 4). Additionally, if the *Planothidium lanceolatum* (PLLA) diatom is more present (more than 6) then the concentration of metals and sulphates (SO_4) is higher.

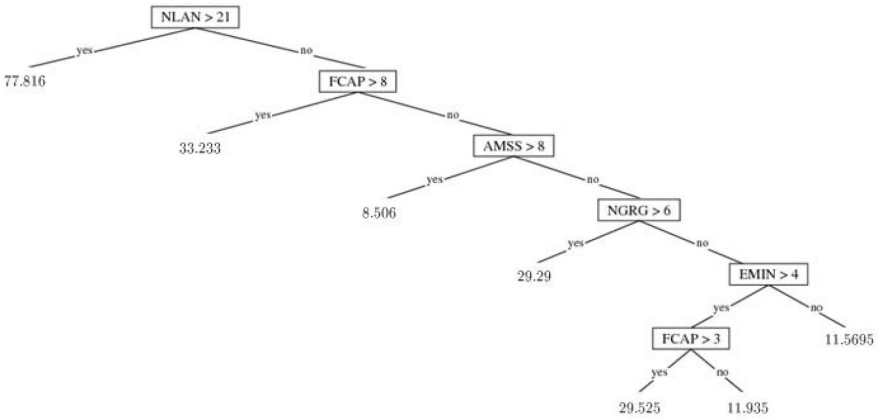

Fig. 5. Regression tree for the total phosphorus from river measurements

The tree presented in Figure 5 shows similar relations between the phosphorus and the diatoms as its more general variant from Figure 4. The highest concentration of phosphorus is expected at NLAN values higher than 21 (as in tree from Figure 4).

If NLAN is present less than 21% (or not present at all), then higher phosphorus concentration are encountered and FCAP is present more than 8%.

6. Conclusion

In this paper, we applied machine learning methodology, in particular regression trees and multiple targets regression trees, to predict the chemical parameters of the environment using the diatom community in Lake Prespa and its tributaries.

Regarding the performance, in our case, RTs [12] achieve slightly better correlation coefficients than MTRTs. However, the presented methodology of multi-target regression trees has several advantages with respect to the more commonly used approach of single target regression trees. Namely, the MTRTs provide knowledge about all targets and, in our case, identify the diatom taxa that are present in the water samples under some specific physic-chemical conditions. On the other hand, using the traditional approach, one would have to construct separate model for each chemical parameter and to summarize over the multiple models, which is not a trivial task.

For further work, we intend to define water classes based on the chemical parameters and construct trees to predict the water classes from the diatom abundance. We, also, intend to acquire data with better quality and obtain trees with better correlation coefficients.

Reference:

[1] Salomoni S, Rocha O, Callegaro V, Lobo E, (2006) Epilithic diatoms as indicators of water quality in the Gravataí River, Rio Grande do Sul, Brasil, Hydrobiologia, 559: 233-246

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

[3] 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

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

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

[6] TRABOREMA Project WP3.3, EC FP6-INCO project no. INCO-CT-2004-509177, pp 98, (2005-2007)

[7] Blockeel H, Struyf J, (2002) *Efficient algorithms for decision tree cross-validation*, Journal of Machine Learning Research 3 (Dec), pp. 621-650

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

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

[10] Stoermer E, Smol J, (eds.) (1999) The Diatoms: Applications for the Environmental and Earth Sciences. Cambridge University Press, Cambridge, 469 pp.

[11] Krstic S, Svircev Z, Levkov Z, Nakov T, (2007) Selecting appropriate bioindicator regarding the WFD guidelines for freshwaters - a Macedonian experience. *International Journal on Algae* 9(1), 41-63.

[12] Dzeroski, S., Mitreski, K., KRSTIĆ, S., Naumoski, A. (2007): Learning habitat models for the diatoms of lake Prespa. V: STANKOVSKI, Mile J. (ed.). ETAI-2007: proceedings of abstracts of VIII National Conference with International participation, Ohrid, Republic of Macedonia, September 19-21, 2007, [COBISS.SI-ID 21291815], sect I1-1, 1-6.

[13] KRSTIĆ S. and Levkov Z. (2007): Saprobiological and trophic models for Lake Prespa (saprographs) for use in similar regions and its application for evaluation of Ecological Quality Ratios (indicators).