Ecological Modelling 221 (2010) 330–337

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Learning habitat models for the diatom community in Lake Prespa

Dragi Kocev[a], Andreja Naumoski[b], Kosta Mitreski[b], Svetislav Krstić[c], Sašo Džeroski[a,*]

[a] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[b] Department of Computer Technologies and Environment Centre, Faculty of Electrical Engineering and Information Technology, Skopje, Macedonia
[c] Institute of Biology, Faculty of Natural Sciences and Mathematics, Skopje, Macedonia

ARTICLE INFO

Article history:
Received 15 May 2008
Received in revised form 26 August 2009
Accepted 4 September 2009
Available online 12 October 2009

Keywords:
Diatom community
Habitat modelling
Multi-target modelling
Regression trees
Lake Prespa

ABSTRACT

Habitat suitability modelling studies the influence of abiotic factors on the abundance or diversity of a given taxonomic group of organisms. In this work, we investigate the effect of the environmental conditions of Lake Prespa (Republic of Macedonia) on diatom communities. The data contain measurements of physical and chemical properties of the environment as well as the relative abundances of 116 diatom taxa. In addition, we create a separate dataset that contains information only about the top 10 most abundant diatoms. We use two machine learning techniques to model the data: regression trees and multi-target regression trees. We learn a regression tree for each taxon separately (from the top 10 most abundant) to identify the environmental conditions that influence the abundance of the given diatom taxon. We learn two multi-target regression trees: one for modelling the complete community and the other for the top 10 most abundant diatoms. The multi-target regression trees approach is able to detect the conditions that affect the structure of a diatom community (as compared to other approaches that can model only a single target variable). We interpret and compare the obtained models. The models present knowledge about the influence of metallic ions and nutrients on the structure of the diatom community, which is consistent with, but further extends existing expert knowledge.

1. Introduction

Ecology is frequently defined as the study of the distributions and abundances of organisms across space and time and their interactions with the environment (Begon et al., 2006). Habitat modelling focuses on the spatial aspects of the distribution and abundance of plants and animals. It studies the relationships between some environmental variables and the presence/abundance of plants and animals. This is typically done under the implicit assumption that both are observed at a single point in time for a given spatial unit.

The input to a habitat model (Džeroski, 2001, 2009) is a set of environmental characteristics for a given spatial unit of analysis. These environmental characteristics (i.e., environmental variables) may be of three different types. The first type concerns abiotic properties of the environment, e.g., physical and chemical characteristic thereof. The second type concerns some biological aspects of the environment, which may be considered as an external impact on the group of organisms under study. Finally, the variables of the third type are related to human activities and their impacts on the environment. The output of a habitat model is a target property of the given (taxonomic) group of organisms. Note that the type of environmental variables, as well as the size of the spatial unit, can vary considerably, depending on the context, and so can the target property of the population (even though to a lesser extent). If we take the abundance or density of the population as indicators of the suitability of the environment for the group of organisms studied, we talk about habitat suitability models: the output of these models can be interpreted as a degree of suitability. The abundance of the population can be measured in terms of the number of individuals or their total size (e.g., the dry biomass of a certain species of algae). If the (taxonomic) group is large enough, we can also consider the diversity of the group (e.g., Shannon index, species richness).

In the most general case of habitat modelling, we are interested in the relation between the environmental variables and the structure of the population at the spatial unit of analysis (absolute and relative abundances of the organisms in the group studied). One approach to this is to build habitat models for each of the organisms (or lower taxonomic units) in the group, then aggregate the outputs of these models to determine the structure of the population. An alternative approach is to build a model that simultaneously predicts the presence/abundance of all organisms in the group.

In this work, we explore the two afore mentioned possibilities for habitat modelling of the diatom community in Lake Prespa

* Corresponding author.
E-mail addresses: Dragi.Kocev@ijs.si (D. Kocev), Andreja.Naumoski@feit.ukim.edu.mk (A. Naumoski), komit@feit.ukim.edu.mk (K. Mitreski), skrstic@iunona.pmf.ukim.edu.mk (S. Krstić), Saso.Dzeroski@ijs.si (S. Džeroski).

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.09.002

(Republic of Macedonia). To learn a model for each diatom taxon separately, we employ regression trees (Breiman et al., 1984). To build a model for the entire diatom community, we use multi-target regression trees (Blockeel et al., 1998; Struyf and Džeroski, 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.

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. The measurements comprise several important parameters that reflect the physical, chemical and biological aspects of the water quality of the lake. These include measurements of the relative abundance of algal 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 environment (Lake Prespa).

Diatoms have narrow tolerance ranges for many environmental variables and respond rapidly to environmental change. This makes them ideal bio-indicators (Reid et al., 1995; Round, 1991). They are sensitive to changes in nutrient concentrations, supply rates and silica/phosphate ratios; they respond rapidly to eutrophication. Each taxon has a specific optimum and tolerance for nutrients such as phosphorus and nitrogen. Diatoms are widely used as bio-indicators in Europe (Krstić, 1995; Krstić et al., 1998; Krstić et al., 2007; Kelly et al., 1998; Prygiel and Coste, 1999), North America (Stevenson and Pan, 1999; Lowe and Pan, 1996), South America (Lobo et al., 1998; Loez and Topalian, 1999) and Australia (John, 1998; Chessman et al., 1999). The geographical location of the diatoms is not the limiting factor in the distribution of diatom taxa and the composition of communities; rather, the specific environmental variables prevailing at a particular location (Gold et al., 2002) are the limiting factors.

The remainder of this paper is organized as follows. In Section 2, we describe the machine learning methodology that was used (regression trees and multi-target regression trees). Section 3 describes the data and Section 4 explains the experimental design that was employed to analyze the data at hand. Section 5 presents the obtained models and discusses them, while Section 6 concludes.

2. Machine learning for habitat modelling

2.1. Machine learning basics

The input to a machine learning algorithm is most commonly a single table of data comprising a number of fields (columns) and records (rows) (Džeroski, 2001, 2009). In general, each row represents an object and each column represents a property (of the object). 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 field from the values of other fields. The target field is called the target attribute or class (dependent variable in statistical terminology). The other fields are called descriptive attributes or just 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 (dataset) 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.

To estimate the performance of the model on unseen data, several approaches can be used (Kohavi, 1995). One approach consists of dividing the data in two parts (typically 2/3 and 1/3): training set (the bigger part) and testing set (smaller part). The most commonly used approach is cross-validation. The division into a training/testing set is recommended in the case of datasets that contain many records (thousands); cross-validation is a better choice otherwise.

2.2. A machine learning formulation of the habitat modelling task

In the case of habitat modelling, examples correspond to spatial units of analysis. The attributes correspond to environmental variables describing the spatial units, as these are the inputs to a habitat model. The class is a target property of the given (taxonomic) group of organisms, such as presence, abundance or diversity.

The machine learning task of habitat modelling (Džeroski, 2009) is thus defined as follows. Given is a set of data with rows corresponding to spatial locations (units of analysis), attributes corresponding to environmental variables, and the class corresponding to a target property of the population studied. The goal is to learn a predictive model that predicts the target property from the environmental variables (from the given dataset). If we are only looking at presence/absence or suitable/unsuitable as values of the class (as is the case above), we have a classification problem. If we are looking at the degree of suitability (density/abundance), we have a regression problem.

2.3. Regression trees

Regression trees are decision trees that are capable of predicting the value of a numeric target variable (Breiman et al., 1984). They 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 predictions for the values of the target attribute are stored in the leaves. Regression tree leaves contain constant values as predictions for the target variable (they represent piece-wise constant functions).

To obtain the prediction of a regression tree 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 procedure is repeated until we end up in a leaf. The resulting prediction of the tree is taken from this leaf.

The tests in the internal nodes can have more than two outcomes (this is usually the case when the test is on a discrete-valued attribute, where a separate branch/subtree is created for each value). Typically, each test has two outcomes: the test has succeeded or the test has failed. The trees in this case are called binary trees.

2.4. Multi-target regression trees

Multi-target regression trees are an instantiation of predictive clustering trees (PCTs) (Blockeel et al., 1998), where a tree is viewed as a hierarchy of clusters. The top-node of a PCT corresponds to a cluster that contains all the data. This cluster is then recursively partitioned into smaller clusters while moving down the tree. The leaves represent the clusters at the lowest level of the hierarchy and each leaf is labelled with its prototype.

Multi-target regression trees (Blockeel et al., 1998; Struyf and Džeroski, 2006) are a generalization of regression trees, because they can predict the values of several numeric target attributes simultaneously. Instead of storing a single numeric value, the leaves

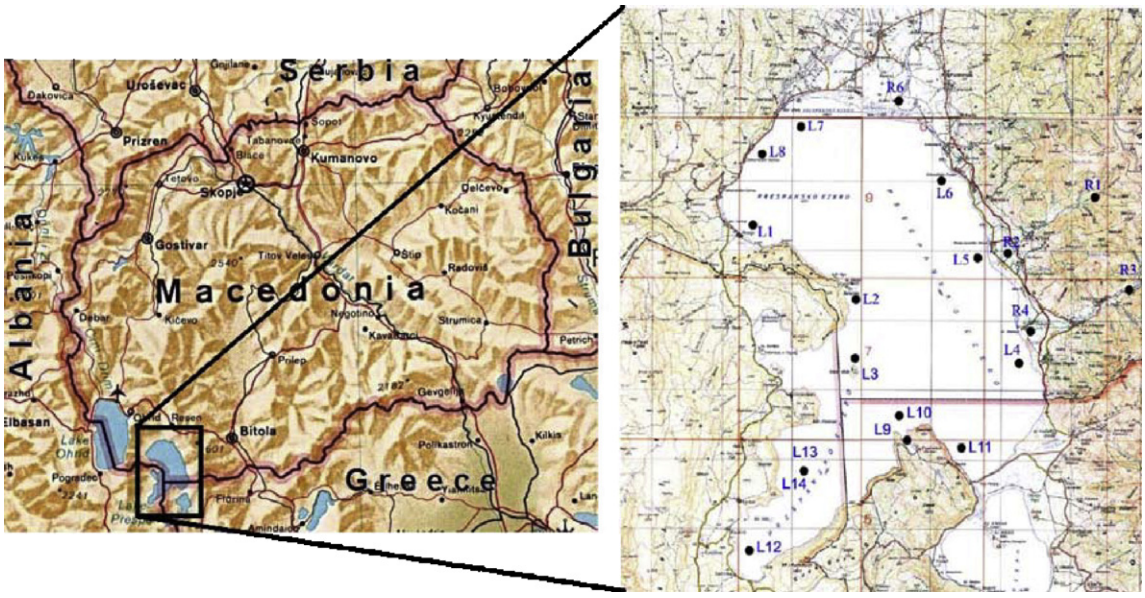

Fig. 1. Position of Lake Prespa (left) and the sampling locations (right).

of a multi-target regression tree store a vector. Each component of this vector is a prediction for one of the target attributes. Examples of multi-target regression trees can be found in Sections 5 and 6.

A multi-target regression tree (of which a regression tree is a special case) is usually constructed by a recursive partitioning algorithm from a training set of records. The algorithm is known as top-down induction of decision trees (TDIDT). The records include measured values of the descriptive and the target attributes. The tests in the internal nodes of the tree refer to the descriptive, while the predicted values in the leaves refer to the target attributes.

The TDIDT algorithm starts by selecting a test for the root node. Based on this test, the training set is partitioned into subsets according to the test outcome. In the case of binary trees, the training set is split into two subsets: one containing the records for which the test succeeds (typically the left subtree) and the other contains the records for which the test fails (typically the right subtree). This procedure is recursively repeated to construct the subtrees.

The partitioning process stops if a stopping criterion is satisfied (e.g., the number of records in the induced subsets is smaller than some predefined value; the depth/size of the tree exceeds some predefined value, etc.). In that case, the prediction vector is calculated and stored in a leaf. The components of the prediction vector are the mean values of the target attributes calculated over the records that are sorted into the leaf.

One of the most important steps in the tree induction algorithm is the test selection procedure. For each node, a test is selected by using a heuristic function computed on the training data. The goal of the heuristic is to guide the algorithm towards small trees with good predictive performance. The multi-target regression trees are implemented in the system CLUS (Blockeel and Struyf, 2002) available at http://www.cs.kuleuven.be/~dtai/clus/. The heuristic used in this algorithm for selecting the attribute tests in the internal nodes is intra-cluster variation summed over the subsets induced by the test. Intra-cluster variation is defined as

$$N \sum_{t=1}^{T} \mathrm{Var}[y_t]$$

with N the number of examples in the cluster, T the number of target variables, and $\mathrm{Var}[y_t]$ the variance of target variable y_t in the cluster. Lower intra-subset variance results in predictions that are more accurate. The variance function is standardized so that the relative contribution of the different targets to the heuristic score is equal.

3. Data description

Lake Prespa is located at the border intersection of Macedonia, Albania and Greece (see Fig. 1). It covers an area of 301 km^2 at 850 m above sea level. The whole region that surrounds the lake was recently proclaimed a transboundary park (Prespa Park). The Prespa Park is well known for its great biodiversity, natural beauty and populations of rare water birds. However, the ecological integrity of the region is threatened by the increasing exploitation of the natural resources (inappropriate water management, forest destruction leading to erosion, overgrazing), inappropriate land-use practices, ecologically unsound irrigation practices, water and soil contamination from uncontrolled use of pesticides, lake siltation and uncontrolled urban development. Monitoring of the state of Lake Prespa is necessary to prevent major catastrophes in the Prespa ecosystem (Krstić, 2006).

Table 1
Basic statistics of the data on physico-chemical water properties obtained from the measurements: minimal value, maximal value, mean value and standard deviation.

	Minimum	Maximum	Mean value	Standard deviation
Temperature (°C)	2.90	26.80	15.56	6.61
Saturated O_2 (mg/dm^3)	6.60	114.19	83.07	19.54
Secchi Depth (m)	1.80	5.40	3.09	0.76
Conductivity (μS/cm)	142.50	318.00	196.23	27.84
pH	5.50	9.27	8.17	0.647
NO_2 (mg/dm^3)	0.00	0.44	0.03	0.05
NO_3 (mg/dm^3)	0.00	13.40	2.07	2.13
NH_4 (mg/dm^3)	0.01	1.07	0.29	0.18
Total N (mg/dm^3)	0.32	9.21	2.53	1.28
Organic N (mg/dm^3)	0.02	8.41	1.83	1.10
SO_4 (mg/dm^3)	2.68	266.10	29.47	22.98
Total P (μg/dm^3)	1.15	83.13	18.63	15.31
Na (mg/dm^3)	0.75	13.15	4.36	2.10
K (mg/dm^3)	0.23	4.80	1.50	0.66
Mg (μg/dm^3)	1.11	19.45	5.70	2.84
Cu (μg/dm^3)	1.04	23.30	3.97	2.79
Mn (μg/dm^3)	0.88	230.00	7.88	16.79
Zn (μg/dm^3)	0.27	22.70	5.23	4.42

Monitoring of the state of Lake Prespa was performed during the EU project TRABOREMA. The measurements cover a one and a half year period (from March 2005 to September 2006). Samples for analysis were taken from the surface water of the lake at 14 locations. The lake sampling locations are distributed in the three countries (see Fig. 1) as follows: eight in Macedonia, three in Albania and three in Greece. The selected sampling locations are representative for determining the eutrophication impact (Krstić, 2005).

Through the lake measurements, a total of 218 water samples were collected. 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, alkalinity (pH), nitrogen compounds (NO_2, NO_3, NH_4, inorganic nitrogen), sulphur oxide ions SO_4, and sodium (Na), potassium (K), magnesium (Mg), copper (Cu), manganese (Mn) and zinc (Zn). The basic statistics for these variables are given in Table 1.

The biological variables were the relative abundances of 116 different diatom taxa (for a complete list of diatom names and acronyms see Table A1 in the Appendix). Diatom cells were collected with a planktonic net or as attached growth on submerged objects (plants, rocks or sand and mud). This is the usual approach in studies for environmental monitoring and screening of diatom abundance. The sample, afterwards, is preserved and the cell content is cleaned. The sample is examined with a microscope, and the diatom taxa and abundance in the samples are obtained by counting 200 cells per sample. The specific taxon abundance is then given as the percent of the total diatom count per sampling site (Levkov et al., 2006).

4. Machine learning experiments and results

4.1. Methodology for constructing models

In this section, we describe the experimental setup used to construct models of the diatom community from the data at hand. The problem we are considering here is the modelling of multiple target variables (responses). As mentioned in Section 1, one approach is to learn a separate model for each target (i.e., diatom taxon) and another one is to learn a single model for all targets (i.e., the complete diatom community).

We analyze the data according to three scenarios: (1) learning a multi-target regression tree for all 116 diatoms (complete community), (2) learning a multi-target regression tree for the top 10 most abundant diatoms and (3) learning regression trees for each diatom separately.

To prevent over-fitting of the models to the training data, we employed '*F*-test pruning'. This pruning method applies the statistical *F*-test (Lomax, 2007) to check whether a given split reduces the variance significantly at a given significance level. The significance level is a user defined parameter: We employ internal 10-fold cross-validation to select an optimal value for this parameter from the following set of values: 0.001, 0.005, 0.01, 0.05, 0.1, 0.125, 0.25, 0.5, 0.75, 1.0. In addition, to obtain even smaller trees, we set a constraint that does not allow the trees to grow more than four levels in depth.

4.2. Predictive power of the models

For each of the learned models, we estimate its predictive performance on both the training data and on unseen data (by 10-fold cross-validation). We use two metrics to evaluate the performance: correlation coefficient and root mean squared error (RMSE). In addition, we inspect the selected models in detail and interpret the knowledge contained therein and compare it to existing knowledge held by a domain expert in the area (S. Krstić).

Table 2
Performance of the regression trees (RT) for predicting the abundance of the top 10 most abundant diatoms on training data and on unseen data (estimated by 10-fold cross-validation).

	F-value	CC		RMSE		Size
		Train	Xval	Train	Xval	
Amphora pediculus (APED)	0.001	0.48	0.18	2.47	2.91	7
Cyclotella juriljii nom. nud. (CJUR)	0.050	0.69	0.12	5.34	8.09	17
Cyclotella ocellata (COCE)	0.001	0.68	0.35	15.70	21.18	21
Cocconeis placentula (CPLA)	0.050	0.78	−0.04	3.14	6.74	21
Cavinula scutelloides (CSCU)	0.001	0.47	0.22	7.75	8.98	9
Diploneis mauleri (DMAU)	0.001	0.42	0.23	2.42	2.69	5
Navicula prespanense (NPRE)	0.001	0.64	0.37	2.17	2.73	11
Navicula rotunda (NROT)	0.001	0.44	0.18	3.15	3.63	13
Navicula subrotundata (NSROT)	0.005	0.65	0.04	3.53	5.26	17
Staurosirella pinnata (STPNN)	0.010	0.69	0.08	2.17	3.49	25

CC: correlation coefficient; RMSE: root mean squared error.

The performance figures for the models learned for the 10 most abundant diatoms are listed in Tables 2 and 3. Each of these tables presents the selected significance level for the *F*-test pruning, the performance (correlation coefficient and RMSE) and the size (total number of nodes, including leaves and internal nodes) of the produced tree. Table 2 presents the performance of the regression trees and Table 3 of the multi-target regression tree.

A quick inspection of the results shows that the prediction problem is very difficult: even on the training data, the performance is low. In order to investigate how much we can improve the predictive performance, we employed ensembles (bagging and random forests) of both regression trees (Breiman, 1996, 2001) and multi-target regression trees (Kocev et al., 2007). It is well known that ensemble methods perform better than individual trees and are amongst the top performing methods for predictive modelling (Caruana and Niculescu-Mizil, 2006). The results are presented in Tables A2 and A3 in the Appendix.

The ensemble models have better predictive performance overall. The best correlation coefficient (on unseen data) is 0.54 (bagging and random forest of regression trees), as compared to 0.37 for the regression trees (the tree for the *Navicula prespanense* – NPRE diatom) and 0.34 for the multi-target regression tree (for *Cyclotella ocellata* – COCE diatom). However, if we inspect the performance of each diatom, we can note that for some cases (e.g., *Navicula subrotundata* – NSROT and *Staurosirella pinnata* – STPNN) the single trees have equal or even slightly better predictive performance than the ensembles.

Table 3
Performance of the multi-target regression tree (MTRT) for predicting the abundance of the top 10 most abundant diatoms on training data and on unseen data (estimated by 10-fold cross-validation).

	F-value	CC		RMSE		Size
		Train	Xval	Train	Xval	
Amphora pediculus (APED)	0.01	0.36	0.18	2.63	2.82	13
Cyclotella juriljii nom. nud. (CJUR)		0.42	0.26	6.73	7.20	
Cyclotella ocellata (COCE)		0.53	0.34	18.23	20.43	
Cocconeis placentula (CPLA)		0.62	0.04	3.95	5.17	
Cavinula scutelloides (CSCU)		0.45	0.27	7.83	8.66	
Diploneis mauleri (DMAU)		0.40	0.15	2.44	2.73	
Navicula prespanense (NPRE)		0.44	0.16	2.55	2.88	
Navicula rotunda (NROT)		0.41	0.24	3.20	3.44	
Navicula subrotundata (NSROT)		0.32	0.17	4.39	4.60	
Staurosirella pinnata (STPNN)		0.24	0.15	2.90	2.97	

CC: correlation coefficient; RMSE: root mean squared error.

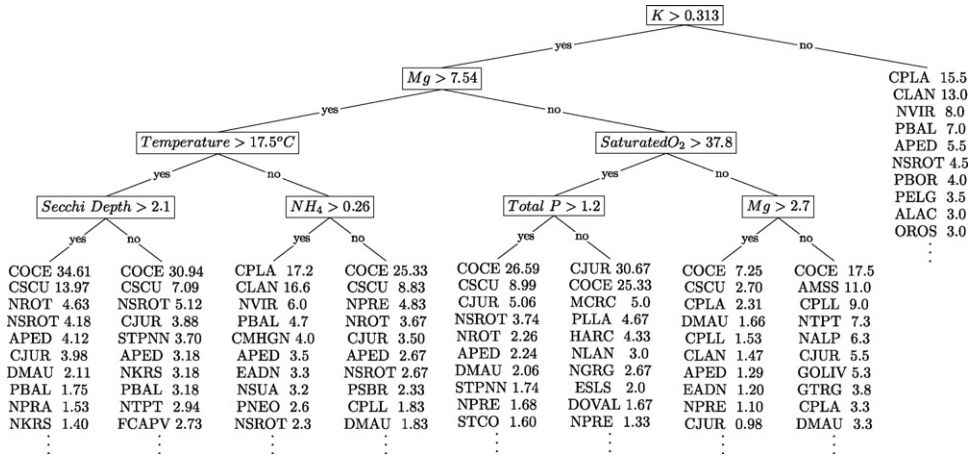

Fig. 2. A multi-target regression tree predicting the structure of the diatom community.

A comparison of the performance of multi-target regression trees and regression trees shows that multi-target regression trees perform better than the regression trees on unseen data. But, on training data, regression trees have better performance. This means that the regression trees tend to over-fit, although the selected significance levels for the *F*-test pruning are quite low (0.001 in the majority of cases – 6/10).

We can also compare the regression trees and the multi-target regression tree by their size (total number of internal nodes and leaves). The size of the multi-target regression tree is 13, while the size of the regression trees ranges from 5 (for *Diploneis mauleri* – DMAU) to 25 (for *S. pinnata* – STPNN), with the 10 trees having a total of 146 nodes. The total size of all single-target trees is much larger that the size of the multi-target tree when we learn a regression tree for each of the 116 diatoms.

In this domain, classical statistical approaches, such as canonical correspondence analysis (CCA), detrended correspondence analysis (DCA) and principal component analysis (PCA), are most widely used as modelling techniques (Stroemer and Smol, 1999). Although these techniques provide useful insights in the data, they are limited in terms of interpretability. On the other hand, multi-target regression trees offer models that are readily interpreted. Also, with these models we are able to identify some environmental conditions that influence the structure of the diatom communities. To summarize, the multi-target regression trees are models that are easily interpretable, with reasonable size and predictive performance.

5. Models of relative abundance of diatom taxa

We applied the methods described in Section 2, according to the methodology described in Section 4, 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 reasonable size (in most cases, the tree size is 9).

5.1. Models for the diatom community

Fig. 2 shows the tree that describes the complete diatom community structure relative to given environmental conditions. It presents nine different diatom communities. The tree has nine leaves/clusters that correspond to different community structures. We can note that the most influential factors for the diatom community are K (potassium) and Mg (magnesium), as well as the temperature and the oxygen. This model defines the environmental conditions (concentration of potassium, magnesium, saturated oxygen, temperature, etc.) under which certain diatom taxa are dominant over the other taxa.

There are two different types of clusters: clusters where *C. ocellata* (COCE) is dominant (the ratio of its abundance to the abundance of the second taxon in the leaf nodes is between 2.4 and 4.5) and clusters where *Cocconeis placentula* (CPLA) and *Cymbella lanceolata* (CLAN) are dominant over the other taxa. When COCE is dominant, it is (in most cases) followed by *Cavinula scutelloides* (CSCU). The

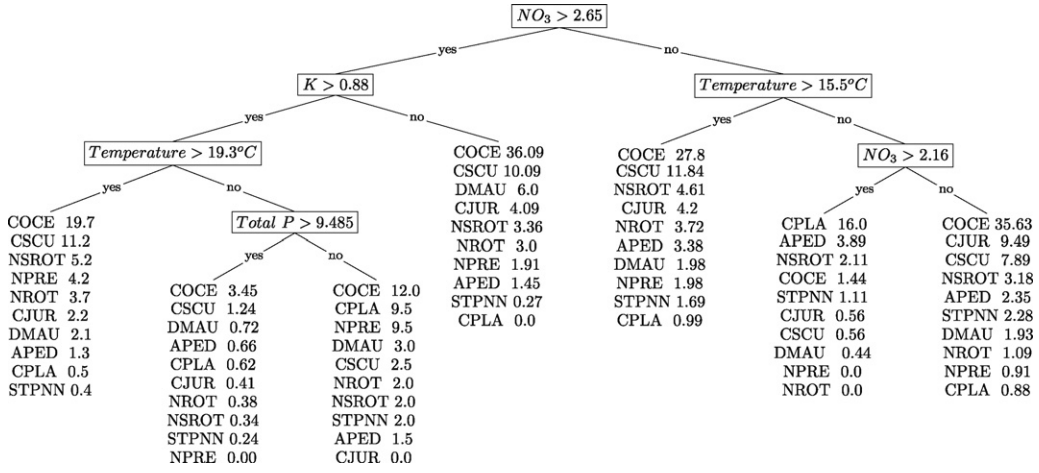

Fig. 3. A multi-target regression tree predicting the relative abundance of the 10 most abundant diatoms.

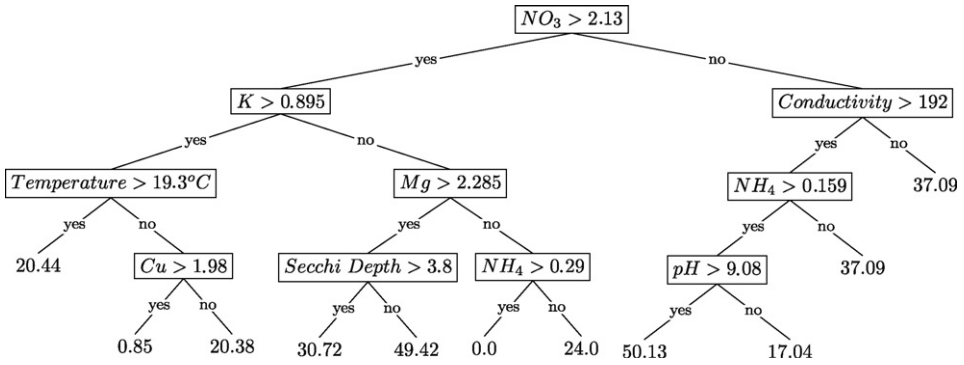

Fig. 4. A regression tree predicting the relative abundance of the diatom *Cyclotella ocellata* (COCE).

Achnanthidium minutissimum (AMSS) and *Cyclotella juriljii* (CJUR) diatoms can also be dominant under some specific environmental conditions. The dominant diatom taxa *C. ocellata* (COCE), *C. scuteloides* (CSCU) and *C. placentula* (CPLA) are all known for their preference towards higher trophic levels (van Dam et al., 1994; Krstić, 2005) and are regarded as precise indicators of high trophic status of the waterbody.

The presented MTRT model of the diatom community structure faithfully reflects the relation of the dominant diatom taxa to the nutrients and physical conditions in the lake. At the top two levels of the tree, we find metallic ions (potassium and magnesium): these are a crucial part of enzymes that play an important role in the life of diatoms. This corresponds to the findings of a previous study from Gold et al. (2002): metal pollution (in particular, cadmium and zinc) affected and changed the diatom community. The model also corresponds to observed diatom flora succession through seasons. The obvious trend of ecological deterioration during summers (the correlation to oxygen content and temperature) is reflected in the increased abundance of eutrophic diatom taxa for Lake Prespa (Krstić and Levkov, 2007).

Fig. 3 depicts the MTRT for the top 10 (most) abundant diatoms in the lake samples. Similar to the model for all diatoms, this model mostly defines clusters where COCE is the dominant taxon (the ratio with the abundance of the second ranked taxon in the leaf nodes is between 1.3 and 3.75). Also, in the majority of these cases (4 out of 6), the second most abundant taxon in the community is the *C. scuteloides* (CSCU) diatom. There is only one cluster where *C. placentula* (CPLA) is dominant.

The model presented in Fig. 3 actually explains more precisely the relations that dominant diatoms in Lake Prespa have with the physico-chemical parameters according to their trophic preferences. Nitrates, as one of the basic external nutrients for the algae and a chemical indicator of higher trophic levels, are the major factor for increasing the relative abundance of *C. ocellata* (COCE); this is yet another piece of evidence of the indicator status of this particular taxon.

The relation of potassium content with the diatom community is more complicated to explain: It may be a result of several factors or their mutual interactions, or even its toxic effects on biota. Thus, the obtained model deserves further attention and investigations, i.e., more broadly conducted research in line with the 'ecosystem approach' that will collect a much more comprehensive database of samples and would consequently yield more precise/accurate models.

Note the difference in the most important factors for predicting the structure of the entire community (Fig. 2) versus the top 10 diatom taxa (Fig. 3). For the top 10 diatom taxa, the nutrients (NO_3) are most important. For the structure of the entire community, however, metal ions play a key role. While the nutrients (nitrogen and phosphorus) are components of proteins, metals such as potassium, magnesium and zinc are parts of (co-)enzymes that also play an important role in cellular processes.

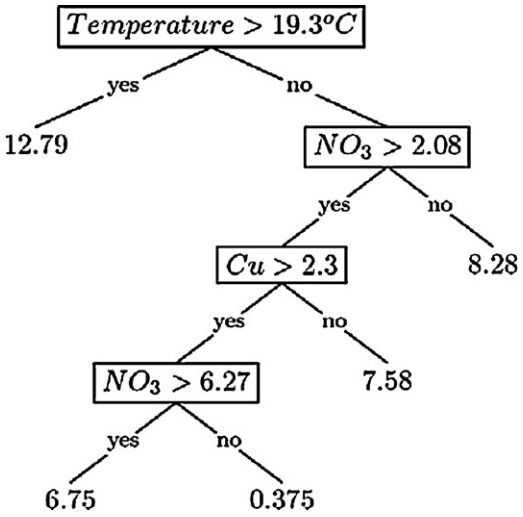

Fig. 5. A regression tree predicting the relative abundance of the diatom *Cavinula scutelloides* (CSCU).

5.2. Models for individual diatoms

We also learned regression trees that predict the relative abundance (habitat suitability) for each of the 10 most abundant diatom taxa separately. We will discuss here the models for *C. ocellata* (COCE, Fig. 4), *C. scutelloides* (CSCU, Fig. 5) and *N. prespanense* (NPRE, Fig. 6): The models for the remaining seven of the top 10 diatoms

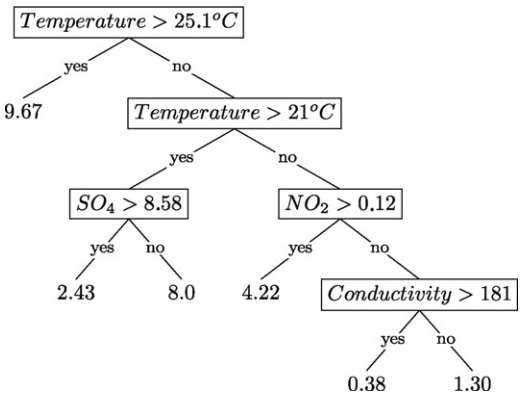

Fig. 6. A regression tree predicting the relative abundance of the diatom *Navicula prespanense* (NPRE).

can be found in Figure A1 in the Appendix. Our choice of the three diatoms to discuss here was motivated as follows: The model for *N. prespanense* (NPRE) is the most accurate one on unseen data, followed closely by *C. ocellata* (COCE), and then with some distance *D. mauleri* (DMAU) and *C. scutelloides* (CSCU) (see Table 2). Of the latter two, we chose *C. scutelloides* (CSCU) as it is the most dominant taxon after *C. ocellata* (COCE).

The most abundant diatom according to the measured data, *C. ocellata* (COCE), is mostly influenced by the nitrogen compounds, the conductivity of the water and the potassium concentration. Other parameters (e.g., NH_4, Cu, Mg, pH, etc., as seen on Fig. 4), specify further where the *C. ocellata* (COCE) diatom is more or less abundant. The absence of *C. ocellata* (COCE) is expected when the concentrations of metals (K and Mg) is low, although there are higher concentrations of nitrogen compounds (NO_3 and NH_4).

The temperature and concentration of nitrates (NO_3) and nitrites (NO_2) are most important for the abundance of the *C. scutelloides* (CSCU) and *N. prespanense* (NPRE) diatoms (see Figs. 5 and 6). These diatoms are most abundant at higher water temperatures (higher than 19.3 °C for *C. scutelloides* – CSCU and 25.1 °C for *N. prespanense* – NPRE). These temperatures are typical for summer periods (especially the one for *N. prespanense* – NPRE). The lowest abundance of CSCU diatoms is encountered at low temperatures, nitrates concentrations between 2.08 and 6.27 and copper concentrations higher than 2.3. The *N. prespanense* (NPRE) diatoms are absent (or present in very small numbers) at lower temperatures (lower than 21 °C), with low concentration of nitrites (NO_2 less than 0.12) and high conductivity of the water (more than 181). In addition, these models identify the limiting role of copper for *C. scutelloides* (CSCU) and of sulphates (SO_4) for *N. prespanense* (NPRE): higher concentrations result in lower abundance.

6. Conclusions

Summary. In this work, we modelled the influence of environmental factors on diatom communities in Lake Prespa. The diatom communities were represented with the relative abundances of the diatom taxa. We applied regression trees and multi-target regression trees to the measured data to model how the structure of diatom communities varies under different environmental conditions.

We first assessed the predictive performance of the obtained models, which were then interpreted for content. The interpretation was done by a domain expert, a biologist who has studied the diatoms in Lake Prespa and collected and processed the samples (S. Krstić). A comparison of the models was then performed along two dimensions. First, we compare the predictive performance of the models both on training data and unseen data. Second, we compare the models by their interpretation in terms of structure and content.

Predictive power. The predictive power of the models on unseen cases is weak (as estimated with 10-fold cross-validation). Since we suspected that over-fitting might play an important role in this, we applied '*F*-test pruning' to prevent over-fitting. However, despite this the predictive power remained poor. On the other hand, the performance on the training data and thus the explanatory power is much better; the tests that are in the nodes produce statistically significant reduction in the variance at a given significance level.

To investigate the limits of predictive performance on the data at hand, we also built ensembles of tree-based models. These are well known for their predictive power and are top performers, at the cost of producing models that are not easy to interpret. This yielded predictive performance that was better than that of a single tree, but still not that high (maximum correlation reached was 0.54).

We can thus conclude that the low predictive performance achieved is not a consequence of using an inappropriate methodology, but rather a consequence of the difficulty of the problem addressed. The modelling problem at hand is very difficult, because the lake is a complex ecosystem and the data available was of limited quantity and quality. In order to obtain models with better predictive power, more measurements are needed. These measurements should include additional locations, a longer period of observation and a wider range of measured environmental parameters.

Model interpretation. Multi-target regression trees are a special case of predictive clustering trees, where the tree is viewed as a hierarchy of clusters. In our study, we focus on the clustering part (how the models describe the training data). All in all, the multi-target regression trees are models that offer easy interpretability and are able to detect the environmental conditions that influence, modify and shape the diatom community as a whole, rather than influence individual taxa.

The developed models clearly reflect and improve the hitherto known ecological preferences of the diatom taxa in Lake Prespa. The dominant lake diatom flora is composed of taxa indicative for increased eutrophication levels and their abundance is directly related to specific physico-chemical parameters. We built models that relate environmental conditions to the relative abundance of the 10 most abundant diatom taxa, as well as to the structure of the entire diatom community.

The models reflect clearly the factors that most influence the abundance of the dominant taxa and the entire community. Metals, nutrients, and temperature are the most important factors for the formation of the community overall. While the nutrients are of key importance for the dominant taxa, it is the metals that most influence the overall community structure.

Conclusion. By using machine learning methods for multi-target prediction, we have improved our understanding of the influence of environmental factors on the diatom community in Lake Prespa. While the learned models do not have strong predictive power, they provide useful explanations that can be related to and can improve upon existing ecological knowledge. The difference of relative importance of environmental factors on the dominant diatoms and the overall diatom community structure is a nice illustration of this.

Multi-target regression trees have been used so far to investigate terrestrial communities, e.g., soil insects (Demšar et al., 2006); to predict chemical parameters of river water quality from bioindicator data (Blockeel et al., 1999) and to predict the condition/quality of indigenous vegetation (Kocev et al., 2009). However, to our knowledge, this is the first use of multi-target regression trees to study lake ecosystems and to investigate/predict the composition of aquatic ecosystem communities.

Future work. In the future, we plan to investigate several research scenarios. One possibility is to take the diatom community as an indicator of water quality and use the diatom abundances as descriptive and the environmental properties as target variables. Another possibility is to represent the diatom community together with its taxonomic structure. The taxonomic structure of the community could then be predicted with hierarchical multi-label classification (Vens et al., 2008) approaches.

Acknowledgement

This work was supported by the bilateral project between Slovenia and Macedonia (Grant number 17/2007-2008), titled "*Knowledge Discovery for Ecological Modelling of Lake Ecosystems*".

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2009.09.002

References

Begon, M., Townsend, C.R., Harper, J.L., 2006. Ecology: From Individuals to Ecosystems, 4th ed. Blackwell, Oxford, England.

Blockeel, H., De Raedt, L., Ramon, J., 1998. Top-down induction of clustering trees. In: Proceedings of Fifteenth International Conference on Machine Learning. Morgan Kaufmann, San Mateo, CA, pp. 55–63.

Blockeel, H., Džeroski, S., Grbović, J., 1999. Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. In: Proceedings of Third European Conference on Principles and Practice of Knowledge Discovery in Databases, LNCS, vol. 1704. Springer, Berlin, pp. 32–40.

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

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the Twenty Third International Conference on Machine Learning, ACM International Conference Proceeding Series 148, New York, NY, pp. 161–168.

Chessman, B., Growns, I., Currey, J., Plunkett-Cole, N., 1999. Predicting diatom communities at the genus level for the rapid biological assessment of rivers. Freshwater Biology 41, 317–331.

Demšar, D., Džeroski, S., Larsen, T., Struyf, J., Axelsen, J., Bruns-Pedersen, M., Henning Krogh, P., 2006. Using multi-objective classification to model communities of soil microarthropods. Ecological Modelling 191 (1), 131–143.

Džeroski, S., 2001. Applications of symbolic machine learning to ecological modelling. Ecological Modelling 146 (1), 263–273.

Džeroski, S., 2009. Machine learning applications in habitat suitability modeling. In: Haupt, S.E., Pasini, A., Marzban, C. (Eds.), Artificial Intelligence Methods in the Environmental Sciences. Springer, Berlin, pp. 397–412.

Gold, C., Feurtet-Mazel, A., Coste, M., Boudou, A., 2002. Field transfer of periphytic diatom communities to assess shortterm structural effects of metals (Cd Zn) in rivers. Water Research 36, 3654–3664.

John, J., 1998. Diatoms: tools for bioassessment of river health. A model for southwestern Australia. In: Land and Water Resources Research and Development Corporation project UCW3, p. 388.

Kelly, M., Cazaubon, A., Coring, E., Dell'Uomo, A., Ector, L., Goldsmith, B., Guasch, H., Hürlimann, J., Jarlman, A., Kawecka, B., Kwandrans, J., Laugaste, R., Lindstrøm, E.-A., Leitao, M., Marvan, P., Padisák, J., Pipp, E., Prygiel, J., Rott, E., Sabater, S., van Dam, H., Vizinet, J., 1998. Recommendations for the routine sampling of diatoms for water quality assessments in Europe. Journal of Applied Phycology 10, 215–224.

Kocev, D., Vens, C., Struyf, J., Džeroski, S., 2007. Ensembles of multi-objective decision trees. In: Proceedings of Eighteenth European Conference on Machine Learning, LNCS, vol. 4701. Springer, Berlin, pp. 624–631.

Kocev, D., Džeroski, S., White, M.D., Newell, G.R., Griffioen, P., 2009. Using single and multi target regression trees and ensembles to model a compound index of vegetation condition. Ecological Modelling 220 (8), 1159–1168.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 1137–1143.

Krstić, S., 1995. Saprobiological characteristics of River Vardar diatom flora as indicator of the intensity of anthropogenic influence. PhD Thesis, Faculty of Natural Sciences and Mathematics, Skopje, Macedonia.

Krstić, S., 2005. Description of sampling sites. Report on baseline data for water (surface and groundwater) including waste related data for the target region. EC-FP6 project "TRABOREMA", EC-Project Contract No. INCO-CT-2004-509177, Deliverable 2.2.

Krstić S., 2006. Report on natural and cultural rates of eutrophication and their impact on microflora biodiversity, past and present spatial and temporal trends and major pressures in the region. EC-FP6 project "TRABOREMA", EC-Project Contract No. INCO-CT-2004-509177, Deliverable 3.2.

Krstić, S., 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). FP6-project TRABOREMA: Deliverable 3.3.

Krstić, S., Levkov, Z., Stojanovski, P., 1998. Diatom communities as indicator of pollution in river Vardar, Macedonia. In: Proceeding of Fifteenth Diatom Symposium, Perth, Australia, Koeltz Scientific Books, Koenigstein, Germany, pp. 103–112.

Krstić, 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.

Levkov, Z., Krstić, S., Metzeltin, D., Nakov, T., 2006. Diatoms of Lakes Prespa and Ohrid (Macedonia) Iconographia Diatomologica 16. Gantner Verlag, Rugell, Lichtenstein.

Lobo, E.A., Callegaro, V.L.M., Bender, E.P., Asai, K., 1998. Water quality assessment of rivers of Southern Brazil using epilithic diatom assemblages. In: Proceedings of Fifteenth International Diatom Symposium, Perth, Australia, Koeltz Scientific Books, Koenigstein, Germany.

Loez, C.R., Topalian, M.L., 1999. Use of algae for monitoring rivers in Argentina with a special emphasis for Reconquista River (region of Buenos Aires). In: Prygiel, J., Whitton, B.A., Bukowska, J. (Eds.), Use of Algae for Monitoring Rivers III, Agence de l'Eau Artois-Picardie, Douai, pp. 72–83.

Lomax, R.G., 2007. Statistical Concepts: A Second Course. Routledge, Oxford, UK.

Lowe, R.L., Pan, Y., 1996. Benthic algal communities as biological indicators. In: Stevenson, R., Bothwell, M., Lowe, R., Thorp, J. (Eds.), Algal Ecology. Freshwater Benthic Ecosystems. Academic Press, San Diego, pp. 705–739.

Prygiel, J., Coste, M., 1999. Progress in the use of diatoms for monitoring rivers in France. In: Prygiel, J., Whitton, B.A., Bukowska, J. (Eds.), Use of Algae for Monitoring Rivers III, Agence de l'Eau Artois-Picardie, Douai, pp. 165–179.

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.

Round, F.E., 1991. Use of diatoms for monitoring rivers. In: Whitton, B.A., Rott, E., Friedrich, G. (Eds.), Use of Algae for Monitoring Rivers. Institut fur Botanik, Universitat Innsbruck, pp. 25–32.

Stevenson, R.J., Pan, Y., 1999. Assessing environmental conditions in rivers and streams with diatoms. In: Stroemer, E.F., Smol, J.P. (Eds.), The Diatoms: Applications for the Environmental and Earth Sciences. Cambridge University Press, Cambridge, pp. 11–40.

Stroemer, E.F., Smol, J.P., 1999. The diatoms: Applications for the Environmental and Earth Sciences. Cambridge University Press, Cambridge.

Struyf, J., Džeroski, S., 2006. Constraint based induction of multi-objective regression trees. In: Proceedings of the Fourth International Workshop on Knowledge Discovery in Inductive Databases, Revised, Selected and Invited Papers, LNCS, vol. 3933. Springer, Berlin, pp. 222–233.

van Dam, H., Mertens, 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.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., Blockeel, H., 2008. Decision trees for hierarchical multi-label classification. Machine Learning 73 (2), 185–214.