Multi-target Modelling of Metal Influence on Diatom Biodiversity – Lake Prespa case study

Andreja Naumoski, Georgina Mirceva, Kosta Mitreski

Ss. Cyril and Methodius University in Skopje, Faculty of Computers Science and Engineering, Skopje, Macedonia
{andreja.naumoski,georgina.mirceva,kosta.mitreski}@finki.ukim.mk

Abstract. The combination of light, temperature, conductivity, oxygen, nitrates and metals form a group of important stress environmental factors that influence on biodiversity. The last ones, the metals, several of these parameters are associated with agricultural and farming activities. Therefore, increased agricultural activity could lead to disruption in biodiversity equilibrium, especially on ecosystem like Lake Prespa. Discovering the right influencing factor on diatom biodiversity is the task that this paper aims to shade a light on. We plan to achieve this by using state of the art methods for machine learning. Since several metal parameters are influencing the diatom biodiversity, multi-target regression tree method is used. We investigate different strategies and we pick the best model(s) based on the experimental evaluation. The obtained models reveal that Na and Mg are the most influencing factors on the diatom biodiversity. Based on these results, further research based on this method for other abiotic stress factors could be made.

Keywords: Multi-Target Regression Models, Over-fitting, Ensembles, Metal Parameters, Biodiversity

1 Introduction

The biodiversity of the lake diatoms is influenced by many factors. Therefore, it is important to know the level of influence of these factors on the biodiversity of one or several organisms that form that community. The biodiversity is mathematical metric that measures the species diversity in the community. The distribution and the abundance of plants and animals is represented using the biodiversity indices, incorporating the value of the spatial aspect. By knowing the properties of the ecosystem biodiversity, we can use methods and techniques to predict certain physico-chemical conditions that favour the particular arrangement of that species distribution. There are methods and techniques in the area of machine learning that build models, so-called biodiversity models, and they try to reveal the relationship between the biological information represented with the biodiversity indices, and the abiotic information known through the physico-chemical parameters information. These models are not possible to be made if the data at hand or the data that we use for modelling is obtained using measurements done in single point in time for a specific geographical area of interest. This kind of

modelling for a moment ignores the temporal aspects of the species distribution, but it can incorporate some temporal aspects that are important for learning the particular model. One example could be the average of the concentration of a particular metal over a time period before the observation.

The biodiversity modelling is very similar to the habitat suitability modelling, but there is a difference. The biodiversity modelling contains "concentrated" biological information of the organisms that is related with the abiotic factors, while the habitat suitability models take into account the abundance of each single specie that compromises that biodiversity. Both type of models relate the biological information with the abiotic parameters. In addition, there are two ways that the biological information can be related with the physico-chemical parameters: building models for each of the abiotic parameters (known as single target models), or to build a model that simultaneously predicts the influence of the entire set of environmental parameters (multi-target models).

In this research paper, we build multi-target biodiversity regression models [1], since the output of the model is numeric value. If the output of the model is nominal, then these models would be multi-class classification models. We use the multi-target regression trees to build models for the metal abiotic factors that influence on the diatom biodiversity, based on the occurrence of a particular diatom in a given time and space [2]. This method has several advantages over single target models, and was used successfully in several studies [3, 4]. His main advantages includes building smaller models, faster learning of the target model and producing interpretable model that reveals the relationship between biodiversity indices and a set of abiotic factors.

The data samples that compromise the data at hand used for obtaining the biodiversity models are collected as a part of the EU funded project TRABOREMA (FP6-INCO-CT-2004-509177) [5]. This data contains information regarding the physico-chemical information and in the same time information regarding the abundance of each collected diatom. The biodiversity information is calculated on the bases of the abundance of the diatom community. In this particular research, we are interested in knowing the relationship only between the set of metal parameters and the biodiversity. This is because the metal parameters previously have shown strong relationship with the diatom community [6].

The rest of this paper is organized as follows. In Section 2, we describe the algorithm for building multi-target regression trees. Section 3 describes the data and then we give a short explanation of the experimental design that was employed to analyse the data at hand. In Section 4, we present the obtained models and discuss them, followed by Section 5 that gives the main conclusions.

2 The Machine Learning Algorithm

Before the models are obtained, the experiment has to be set up. The input of the machine learning method, in the case of biodiversity modelling of the diatom community,

is a set of metal abiotic factors. Therefore, the attributes correspond to the environmental variables describing the ecosystem conditions, and the output of the model is a vector of biodiversity indices.

Accordingly, the machine learning task that we are doing is defined as follows. Input a set of data that the attributes of this dataset corresponds with the metal environmental variables, rows that corresponding to the spatial locations, and the output or the class attribute corresponds the target biodiversity of the diatom community. Therefore, the goal is to learn a predictive model that will reveal the target property from the metal abiotic factors. Since, we are looking for a degree of biodiversity and this is a numeric value, we are solving a regression problem.

2.1 Multi-Target Regression Trees

The method used to construct the biodiversity models or the multi-target regression tree is a generalization of the regression tree method. This is because the former can predict a value of multiple numeric attributes at one compare to the later [1]. Therefore, instead obtaining models with single prediction value, the used method is able to obtain model that predicts several values at once stored in a vector. Each vector element corresponds to a prediction of the target attributes.

The model in a form of a tree is constructed [7] with a recursive partitioning algorithm from a set of records. Usually the records are comprised from the measured values of the descriptive and target attributes. There are two parts in the process of learning the model; training and test selection procedure. These two selection procedures are using separate portions of the dataset. There are different strategies to separate the dataset, but one approach is to take 2/3 of the input dataset and use to train (learn) the model, and 1/3 of the input data to test the model to see how the model is behaving on unseen data. The test procedure is the most important step in the induction algorithm. This procedure includes heuristic function that is computed on the training data and selects each test for a given node. Using this function, the induction algorithms tries to get smaller trees with good predictive performance and reduces the model complexity. In order to construct multi-target regression biodiversity models, we use the CLUS software [8]. CLUS uses sum of inter-cluster variations [9] in the process of building the model as heuristic function for selection of the test.

This allow the method to learn models with more accurate predictions. Additionally, the multi-target regression tree can be pruned with various metrics to improve prediction accuracy, interpretability and reduce model complexity. Several different strategies are imposed on the model in this paper; maximum number of nodes (*maxsize*), maximum depth of the tree (*maxdepth*) and minimal records in a leaf (*minleaf*). Furthermore, in order to reduce the over-fitting of the models, we employ 'F-test pruning'. This pruning uses the statistical F-test [10] to check whether a given split reduces the variance significantly at a given significance level. This is done by making internal 10-fold cross-validation to select an optimal value for this parameter from a set of values.

3 Data Description and Experiments

The data used for learning multi-target biodiversity models were collected during the EU project TRABOREMA and cover period of 16 months. Measurement policy was conducted on both lake and river sampling locations, and in total 275 water samples were acquired. From these measurements, 218 samples represent the lake water chemistry and diatoms' abundances, while 57 reflect the river condition. The biologist conducted both physico-chemical and biological analyses on these samples.

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 Potassium (K), Magnesium (Mg), Sodium (Na), Copper (Cu), Zinc (Zn) and Mangan (Mn).

The biological information was contained in the relative abundances of the 116 different diatoms. For further details of the diatom collection procedure, the reader can refer to [11]. After collection, the samples were examined under microscope, and the number of diatom species is counted. Then the specific species abundances were stored in the database as a percent of the total diatom count per sample [12]. The diatoms' abundances were converted and characterized with 9 biodiversity indices (Chao Richness, Hill N1, Hill N2, 1/Berger_Parker, 1/Simpson, 1/NewSimpson, Shannon, Brillouin, Margalef). In Table 1, the basic statistical information regarding the collected parameters and the biodiversity indices are given.

Table 1. Basic statistics of the data on physico-chemical water properties obtained from the measurements: minimal value (Min), maximal value (Max), mean value (Avg) and standard deviation (Sd) for both lake and rivers datasets, seperatly.

	Lake				Rivers			
Abiotic and Biodiversity indices	Min	Max	Avg	Sd	Min	Max	Avg	Sd
Na	0.75	13.15	4.36	2.1	0.71	8.89	2.09	1.32
K	0.23	4.8	1.5	0.6	0.31	6.65	1.19	1.04
Mg	1.11	19.45	5.7	2.8	0.22	9.63	2.5	2.5
Cu	1.04	23.3	3.97	2.8	0.64	13.28	4.43	3
Mn	0.87	230	7.88	16.8	1.04	79.3	16.51	19.25
Zn	0.27	227.7	5.23	4.4	0.25	214.5	9.84	29.48
Chao Richness	0	63.25	22.06	13.2	0	51	26.25	11.67
Hill N1	1	25.17	9.62	5.76	1	31.52	15.4	7.14
Hill N2	0	20.57	6.25	4.58	0	29.58	11.84	6.26
1/Berger_Parker	0	11.11	3.08	2.11	0	16.67	5.78	3.22
1/Simpson	0	33.15	6.94	5.6	0	41.6	13.96	8.17
1/NewSimpson	0	20.57	6.25	4.58	0	29.58	11.84	6.26
Shannon	0	3.22	1.97	0.91	0	3.45	2.5	0.92
Brillouin	0	2.78	1.74	0.8	0	2.99	2.2	0.8
Margalef	0	7.5	3.55	1.76	0	7.22	4.49	1.9

3.1 Experimental Design

We learn two different type of models: (1) models that are based on the biodiversity information contained in the entire 116 diatoms in the community for lake stations, and (2) models based on biodiversity information in the diatom community for river stations. In both scenarios (lake and river datasets), the biodiversity indices are related with the metal abiotic factors.

In order to increase the prediction accuracy of each model, to reduce model complexity and over-fitting, we applied four different strategies: *minleaf*, *maxdepth*, *maxsize* and *F-test*. For *minleaf* we set the values of 2, 4, 8, 16 and 32; for *maxdepth* we set 3, 4, and 5, while for *maxsize* we set 7, 9, 11 and 13 [13]. For each model tree, the following set of values was used for the *F-test*: 0.05, 0.075, 0.1, 0.125, 0.25, 0.5, 0.75, 1.0. From all these models, we select the ones that have better predictive power, complexity and lowest over-fitting.

Finally, the test procedure that estimates the performance of the prediction model was assessed by using 10-fold cross validation. For comparison of the model predictive power, we used two different metrics: correlation coefficient (CC) and relative root mean squared error (RRMSE).

3.2 Experimental Results

We estimated the predictive power of every learned model for both training and testing data. Additionally, we compute and compare the CC and RMSE of all models obtained for each pruning strategy. In this paper, we present two models (for each dataset) that obtained best predictive performance results. For both lake and river measurements, the best pruning strategy was – *minleaf* 16.

Table 2. Performance of multi-target regression tree (MTRT) for revealing the relationship between the biodiversity indices and the metal parameters on training data and unseen data using lake measurement data. Bolded results show the maximum accuracy based on CC and underlined results shows minimum error for RRMSE for each test.

Biodiversity Indices	MTRT				MTRT Random Forest			
	CC		RRMSE		CC		RRMSE	
	Train	Test	Train	Test	Train	Test	Train	Test
Chao Richness	0.48	0.29	0.880	0.967	0.71	**0.28**	0.707	<u>1.024</u>
Hill N1	0.45	0.20	0.893	1.001	0.79	0.13	0.608	1.156
Hill N2	0.40	0.16	0.917	1.009	0.77	0.08	0.633	1.186
1/BergerParker	0.41	0.19	0.914	0.998	0.76	0.08	0.655	1.175
1/Simpson	0.37	0.14	0.929	1.008	0.76	0.05	0.655	1.195
1/NewSimpson	0.40	0.16	0.917	1.009	0.77	0.08	0.633	1.186
Shannon	**0.60**	0.36	0.801	<u>0.947</u>	**0.86**	0.26	<u>0.516</u>	1.092
Brillouin	**0.60**	**0.37**	<u>0.798</u>	<u>0.947</u>	**0.86**	0.27	<u>0.516</u>	1.091
Margalef	0.58	0.35	0.818	0.950	0.83	0.27	0.552	1.076

Both tables (Table 2 for lake measurements and Table 3 for river dataset), presents the results from these evaluation experiments of the models based on these two metrics (CC and RRMSE). The experimental results in both tables show that finding the true relationship between the biodiversity indices and the metal abiotic factors is not very easy task. The performance of the models obtained on training data is relatively medium, but if we compare the datasets, the lake dataset is more promising. Furthermore, if we inspect the results from the prediction (test) results, it is obvious that some of the biodiversity indices do not correlate with the metal abiotic factors.

Table 3. Performance of multi-target regression tree (MTRT) for revealing the relationship between the biodiversity indices and the metal parameters on training data and unseen data using river measurement data. Bolded results show the maximum accuracy based on CC and underlined results shows minimum error for RRMSE for each test.

Biodiversity Indices	MTRT				MTRT Random Forest			
	CC		RRMSE		CC		RRMSE	
	Train	Test	Train	Test	Train	Test	Train	Test
Chao Richness	0.13	0.08	0.991	1.008	0.52	0.02	0.854	1.093
Hill N1	0.30	0.00	0.953	1.024	0.77	0.08	0.637	1.096
Hill N2	0.30	0.03	0.956	1.030	0.80	0.08	0.598	1.102
1/BergerParker	0.21	0.03	0.977	1.024	0.77	**0.10**	0.637	1.081
1/Simpson	**0.30**	0.01	0.955	1.029	**0.82**	0.06	0.571	1.113
1/NewSimpson	**0.30**	0.03	0.956	1.030	0.80	0.08	0.598	1.102
Shannon	0.12	**0.22**	0.993	1.021	0.63	0.00	0.775	1.111
Brillouin	0.11	**0.25**	0.994	1.020	0.63	0.00	0.779	1.111
Margalef	0.24	0.01	0.970	1.014	0.67	0.04	0.747	1.097

In order to prevent over-fitting of the predictive models, we set the F-test pruning strategy starting from 0.05. This stopped the models to have large difference between the training and test performance. Most of the models obtained values between 0.05 and 0.1. We also performed experiments in order to know how much we can improve the predictive performance, so we made tests using one ensemble method (random forest) that is one of the top performing methods for predictive modelling [14]. Overall, the descriptive models (models that are obtained using training data) obtained using the random forest method have better performance for both datasets, but the predictive performance for both datasets are lower than the predictive performance of multi-target tree.

4 Biodiversity Models

We used the methodology described in Section 2, based on the description of the experimental setup in Section 3.1, and we used the biodiversity data that contain information regarding the biodiversity indices and the metal abiotic factors for both lake and river sites. During the process of learning the models, we used the settings described in

previous section and we obtained two multi-target regression tree models for each dataset with best predictive performance. Both models were learned using *minleaf* 16 strategy. In this way, we obtained two predictive models that describe the influence of the set of metal measured parameters on all biodiversity indices taken into account.

4.1 Models for the Lake Measurements

Fig. 1 presents a predictive tree model that relates the biodiversity indices relative to given metal environmental condition. In the nodes of the tree, we can distinct four different clusters that correspond to four different biodiversity scenarios. The most obvious is the root of the tree that depicts the most influencing factor on the biodiversity indices: Na (sodium) and Mg (magnesium) as well as the K (potassium) component.

From the four biodiversity scenarios, two different clusters emerging from the model; one cluster where the biodiversity indices have very similar values and have much higher results than the second cluster, where biodiversity indices have very low values. Table 1 gives us a reference point from which we can estimate the suitable conditions for the lake diatom biodiversity. Consequently, it is obvious that the third cluster from left to right is the cluster that depicts the environmental conditions that are not in favour of diatom biodiversity. For values of sodium larger than 5.2 mg/dm^3, magnesium lower than 7.28 mg/dm^3 and potassium lower than 1.605 mg/dm^3, the diatom biodiversity is very low for all indices. This is very important result from the model, because a set of metal abiotic conditions was found under which all the biodiversity indicators have low or high values.

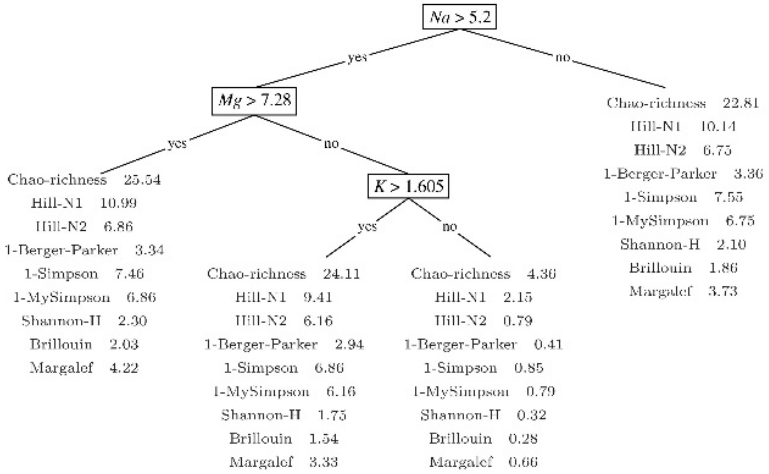

Fig. 1. Multi-target regression tree obtained using lake station measured data with *minleaf* 16 pruning strategy

The results from the presented model realistically depicts the relationship that metals have with the diatom biodiversity in Lake Prespa. According to [6], the most important environmental factors on the diatom community are the metal parameters – Cu, Mn and K. The explanation of the interconnection between these parameters in the diatom community is more complicate to explain. Due to the complexity of the interaction of the nutrients, diatoms and the influencing factors it is difficult to give single conclusion based on several models. Therefore, the obtained model deserves further attention and investigation, i.e., more broadly conducted research on these relationships.

4.2 Models for the Rivers Measurements

Using a similar approach as for the lake measurement model, the multi-target regression tree model for the river measurements depicts the influence of the river chemistry on the diatom biodiversity. Fig. 2 depicts the influence of a single metal abiotic factor on the biodiversity indices.

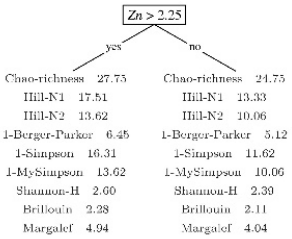

Fig. 2. Multi-target regression tree obtained using lake station measured data with *minleaf* 16 pruning strategy

Compared with the model presented in Fig.1, here the two clusters/leafs of the model do not find any big difference of the biodiversity indices values. Therefore, it is very difficult to distinct the degree on which the value of influence factor, in this case the Zn (zinc), is influencing the river diatom biodiversity. Although it is very interesting to compare the findings in [14], where the Zn parameters influence the biodiversity of the entire lake. The models presented in [14] does not include the data from the river measurements. This could imply that the Zn influencing factor found in [14] could have source from the river tributaries. Thus, the obtained models deserve further investigation and attention.

5 Conclusion

In this paper, we learned several environmental models using machine learning methodology, in particular multi-target regression trees, that model the metal abiotic influence of the diatom biodiversity in Lake Prespa and its tributaries. We modelled biodiversity indices from two diatom communities (one found in lake and one in rivers) that have different structure and different environmental preferences.

If we observe the obtained multi-target regression trees, we can recognise the conditions that increase or decrease the biodiversity of the diatom community, and we can relate them with the biodiversity indices of other measurements (lake vs rivers), and then we can see how the studied objects influence each other. Furthermore, using the multi-target regression trees, we identified for lake measurements conditions in which biodiversity is low or high. Someone can criticize these biodiversity models on a ground that these models do not include all the information from the interaction of the diatoms itself. Yet, the models successfully completed their mission and were in line with the known ecological preferences of the diatom tax found in Lake Prespa. Even the research area is complex, the multi-target regression tree was able to give an inside look of the complex interlocked relationships between the diatom biodiversity and the metal abiotic factors. For further work, we plan to model the diatom biodiversity indices with some other abiotic parameters that are important for the diversity of the community.

Acknowledgement: This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

6 References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth (1984)
2. Levkov Z., Krstic S., Metzeltin, D and Nakov, T.: Diatoms of Lakes Prespa and Ohrid (Macedonia). Iconographia Diatomologica Vol. 16, 603 (2006)
3. Kocev, D., Džeroski, S., White, M. D., Newell, G. R., Griffioen, P.: Using single and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecological Modelling, Vol. 220, no.8, 1159-1168 (2009)
4. Kocev, D., Naumoski, A., Mitreski, K., Krstic, S., Dzeroski, S.: Learning habitat models for the diatom community in Lake Prespa, Ecological Modelling, 221(2), pp. 330-337, 2010.
5. TRABOREMA Project WP3, EC FP6-INCO project no. INCO-CT-2004-509177, 2005-2007
6. Levkov, Z., Blanco, S., Krstic, S., Nakov, T., Ector, L.: Ecology of benthic diatoms from Lake Macro Prespa (Macedonia). Algological Studies, Vol. 124, no.1, 71-83 (2007)
7. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: Shavlik, J. (Ed.), Proceedings of the 15th International Conference on Machine Learning, 55–63, (1998)
8. Blockeel, H., Struyf, J.: Efficient algorithms for decision tree cross-validation, Journal of Machine Learning Research, Vol. 3, 621-650 (2002)
9. Struyf, J., Dzeroski, S.: Constraint based induction of multi-objective regression trees, Knowledge Discovery in Inductive Databases, 4th International Workshop, KDID'05, LNCS vol. 3933, 222-233 (2006)
10. Lomax, R.G.: Statistical Concepts: A Second Course. Routledge, Oxford, UK (2007)
11. WFD Water Quality - Sampling - Part 2: Guidance on sampling techniques (ISO 5667-2:1991) (1993)
12. TRABOREMA Project WP3, EC FP6-INCO project no. INCO-CT-2004-509177, (2005-2007)

13. Caruana, R., Niculescu-Mizil, A.: 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, 161–168 (2006)
14. Naumoski, A.: Multi-target modelling of diatoms diversity indices in Lake Prespa. Applied Ecology and Environmental Research Vol. 10, no.4, 521-529 (2012)