Relating Biodiversity of River Communities to Physical and Chemical Water Properties

Sašo Džeroski[1] and Jasna Grbović[2]

Abstract

We address the problem of finding relationships between the physical and chemical properties of river water and the biodiversity of the community present in that water. We apply the machine learning approach of induction of regression trees to biological and chemical data collected through regular monitoring of rivers in Slovenia. A predictive model is built, which identifies the most important parameters for predicting the species richness (the number of taxa) of the community: these include biological oxygen demand (an overall indicator of pollution), water temperature, the season (month), total hardness, NO_3, SiO_2 and alkalinity.

1. The data

In this study, we use biological and chemical data collected through regular monitoring of rivers in Slovenia. The data come from the Hydrometeorological Institute of Slovenia (Hidrometereološki Zavod Republike Slovenije, abbreviated as HMZ) that performs water quality monitoring for most Slovenian rivers and maintains a database of water quality samples. The data provided by HMZ cover a six year period, from 1990 to 1995.

Biological samples are taken twice a year, once in summer and once in winter, while physical and chemical analyses are performed several times a year for each sampling site. The physical and chemical samples include the measured values of 15 different parameters: biological oxygen demand (BOD), chlorine concentration (Cl), CO_2 concentration, electrical conductivity, chemical oxygen demand ($K_2Cr_2O_7$ and $KMnO_4$), concentrations of ammonia (NH_4), NO_2, NO_3 and dissolved oxygen (O_2), alkalinity (pH), PO_4, SiO_2, water temperature, and total hardness.

The biological samples include a list of all taxa present at the sampling site and their density. The taxa are identified mostly at the species level, so a sample might state that *Tubifex tubifex* was present. Sometimes, however, taxa might be identified to the genus level only (*Tubifex sp.*) or even the family level only (*Tubificidae*).

[1]Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia,
email: Saso.Dzeroski@ijs.si
[2]Hydrometeorological Institute, Vojkova 1b, SI-1000 Ljubljana, Slovenia,
email: Jasna.Grbovic@rzs-hm.si

In total, 1060 water samples were available on which both physical/chemical and biological analyses were performed. The experiments presented were conducted using these samples. These data have been previously used to learn habitat suitability models, i.e., identify the ecological requirements, for selected taxa (Džeroski et al. 1998).

2. Experimental setup

2.1 The task

The particular task addressed was the prediction of the number of taxa present in a biological sample. The number of taxa is a measure of biodiversity. As taxa are mostly identified to species level, it is closely related to species richness.

The number of taxa is the dependent variable that we have to predict. The physical and chemical water properties in the corresponding chemical sample are taken as independent variables. We also added the month when the sample was taken to the independent variables, as the season may have considerable influence on the number of taxa found at a sampling site.

The goal of this study was to use data collected through monitoring Slovenian rivers to build a model. The model can be then used to predict the number of taxa that we can expect at a site from the physical and chemical properties of the river water at that site. It can also be used to provide some insight into which of the the physical and chemical properties of the river water are most related to the number of taxa.

2.2 Regression trees

We approached the above task using the machine learning approach of regression tree induction (Breiman et al. 1984). Regression trees are a representation for piece-wise constant or piece-wise linear functions. They predict the value of a dependent variable (called class) from the values of a set of independent variables (called attributes).

Regression trees are learnt (constructed automatically) from data represented in the form of a table. Columns in the table correspond to measured parameters, while rows correspond to samples. Each row (example) has the form (x_1, x_2, ..., x_N, y), where x_i are values of the N independent variables, also called attributes (i.e., BOD, Cl, ...), and y is the value of the dependent variable, also called class (e.g., the number of taxa).

Unlike classical regression approaches, which find one single equation for a given set of data, induction of regression trees partitions the space of examples into axis-parallel rectangles and fits a model to each of these partitions: the model can be a linear one or just a constant. An example regression tree is given in Figure 1. Internal nodes (ovals) contain attributes and branches are labeled with conditions on attribute values. Leaves (rectangles) contain models that predict the value of the class.

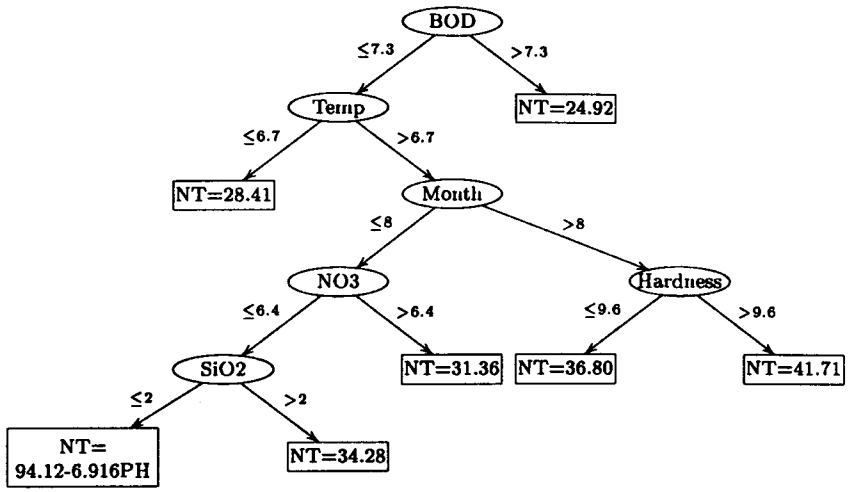

Figure 1: A regression tree predicting the number of taxa (NT) from physical and chemical properties of river water.

To predict the class value for a new sample (example), its attribute values are matched to the conditions in the tree, starting with the root. Suppose we are given a sample with BOD value 8.7. The number of taxa predicted (NT) will be 25. The value of BOD (which is at the root of the tree) is compared to 7.3; the right branch is taken, as 8.7 is greater than 7.3. In this fashion we arrive at the leaf which predicts NT=24.92.

2.3 Methodology

In our study, we built models with the program M5 (Quinlan 1993), one of the most well-known programs for regression tree induction. M5 was used to induce regression trees for predicting the number of taxa in a sample from the values of physical and chemical water properties.

Two regression trees were induced. One was learnt from the entire set of 1060 samples. This tree is given in Figure 1. This tree was inspected to gain insight into the relative importance/influence/correlation of the individual parameters for/on/with biodiversity. Another tree was built to evaluate the predictive performance of regression trees on the task at hand. We took one half of the data for learning (from 1990 to 1992 and some of 1993) and the remaining half for testing.

In both cases, we imposed a minimum number of samples for each leaf in order to ensure the generality of the generated models. This number was set to 100 when learning from the entire dataset and to 50 when learning from one half of the dataset.

3. Results

The regression tree learnt from the entire set of samples is given in Figure 1. We examine this tree in some detail, in order to gain insight into the relative importance of the physical and chemical water properties for biodiversity.

At the top of the tree we find BOD (biological oxygen demand), which seems to influence the number of taxa the most. A large value of BOD (> 7.3) implies about 25 taxa (the average of 115 samples), which is much lower than the overall average (the number of taxa in the biological samples ranges from 6 to 80, with an overall average of 34 taxa). This is understandable, as large BOD values mean heavily polluted water and low water quality, which in turn is correlated with low biodiversity.

If BOD is not large, the number of taxa is lowest in cold water (Temp $\leq$ 6.7), where it is still lower than the overall average (NT=28.41, average of 142 samples). The highest number of taxa is predicted for relatively clean and warm water in the autumn (Month > 8): it is higher in hard waters with Hardness > 9.6 (NT=41.71, average of 136 samples) and lower in softer waters with Hardness ≤ 9.6 (NT=36.80, average of 100 samples). In January to August, the number of taxa is highest in waters which are not too rich with nutrients ($NO_3 \leq 6.4$, $SiO_2 \leq 2$, average NT for 215 samples is 37.92), where it increases as alkalinity decreases (NT $= 94.120 - 6.916$ pH).

To estimate the predictive performance of regression trees on the task at hand, we took one half of the data for learning and the remaining half for testing. The predictive performance of regression trees is slightly better than that of standard linear regression, but both are very low (below 0.3). This is not surprising, since the community present at a river site is not completely determined by the physical and chemical properties of river water at that site at a given moment, but rather depend on the values of these properties over a larger period of time at the given and upstream sites.

The relationships identified by the regression tree on the entire dataset, however, are still interesting, understandable and meaningful to river biologists and water quality experts. They shed some light on the dependence of biodiversity of a community at a river site on the physical and chemical properties of the water at that site. In particular, BOD and water temperature are found to distinguish best among samples of different biodiversity.

4. Related work

An increasing number of studies relating biodiversity to environmental factors is now appearing in the scientific literature. The biodiversity of different groups of organisms is studied. The groups of plants studied include trees (Austin et al. 1996), neotropical trees and lianas (Clinebell et al. 1995) and aquatic macrophytes (Vestergaard/Sand-Jensen 2000).

Many diverse groups of animals have been studied, ranging from crustacean zooplankton (Dodson 1992) through insects (Lek-Ang et al. 1999) to amphibians (Hecnar/McCloskey 1996). Most effort has gone into modelling the relationships between environmental factors and the biodiversity of freshwater macroinvertebrate communities (Lounaci et al. 2000; Brosse et al. 2001) and fish communities (Matrorillo et al. 1998; Gozlan et al. 1999; Lim et al. 1999; Brosse et al. 2001), mostly in rivers. These studies typically use statistical methods (regression) or neural networks to build models for predicting biodiversity from environmental factors (such as altitude, slope, and distance from source for rivers).

5. Discussion

We have addressed the problem of finding relationships between the physical and chemical properties of river water and the biodiversity of the community present in that water. We have used the machine learning approach of induction of regression trees to build a predictive model for the number of taxa from biological and chemical properties of river water. While the predictive power of the model is not high, it identifies the most important parameters for predicting the biodiversity of the community: these include biological oxygen demand (an overall indicator of pollution), water temperature, the season (month), total hardness, NO_3, SiO_2 and alkalinity.

While other studies in relating biodiversity to environmental factors typically use statistical methods (regression) or neural networks to build models, we use the machine learning method of inducing regression trees. Neural networks are black-box models that are not easy to inspect, while regression trees have an explicit and comprehensible structure. This has enabled us to identify the most important parameters for predicting the biodiversity of river communities and has given the built model considerable value, despite its low predictive power.

While most existing studies focus on a relatively small group of organisms (e.g., fish and macroinvertebrates), our study takes all plant and animal taxa into account. While existing studies of biodiversity in rivers typically use environmental factors, our study uses physical and chemical properties of river water. The latter points a direction for further work: taking into account environmental factors such as river-bed composition, altitude, slope, and distance from source should improve the regression trees for predicting biodiversity, both in terms of performance and understandability.

Bibliography

Austin, M.P, (et al.) (1996): Patterns of tree species richness in relation to environment in South-Eastern New South Wales, Australia. *Australian Journal of Ecology*, 21(2): 154–164.

Breiman, (et al.) (1984): *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Brosse, S. (et al.) (2001): Abundance, diversity, and structure of freshwater invertebrates and fish communities: An artificial neural network approach. *New Zealand Journal of Marine and Freshwater Research*, 35(1): 135–145.

Clinebell, R.R. (et al.) (1995): Prediction of neotropical tree and liana species richness from soil and climatic data. *Biodiversity and Conservation*, 4(1): 56–90.

Dodson, S. (1992): Predicting crustacean zooplankton species richness. *Limnology and Oceanography*, 37(4): 848–856.

Džeroski, S., Grbović, J., and Walley, W. J. (1998): Machine learning applications in biological classification of river water quality. In Michalski, R.S., Bratko, I., and Kubat, M., editors, *Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications*, pages 429–448. John Wiley and Sons, Chichester.

Gozlan, R.E., (et al.) (1999): Predicting the structure and diversity of young-of-the-year fish assemblages in large rivers. *Freshwater Biology*, 41(4): 809–820.

Hecnar, S.J., McCloskey, R.T. (1996): Amphibian species richness and distribution in relation to pond water chemistry in South-Western Ontario, Canada. *Freshwater Biology*, 36(1): 7–15.

Lek-Ang, S., Deharveng, L., and Lek, S. (1999): Predictive models of collembolan diversity and abundance in a riparian habitat. *Ecological Modelling*, 120(2–3): 247–260.

Lounaci, A., (et al.) (2000): Abundance, diversity and community structure of macroinvertebrates in an Algerian stream: The Sebaou wadi. *Annales de Limnologie – International Journal of Limnology*, 36(2): 123–133.

Lim, P., (et al.) (1999): Diversity and spatial distribution of freshwater fish in Great Lake and Tonle Sap river, Cambodia. *Aquatic Living Resources*, 12(6): 379–386.

Mastrorillo, S., (et al.) (1998): Predicting local fish species richness in the Garonne River basin. *Comptes Rendus de l Academie des Sciences Serie III – Sciences de la Vie – Life Sciences*, 321(5): 423–428.

Quinlan, J. R. (1993): Combining instance-based and model-based learning. In *Proc. Tenth International Conference on Machine Learning*, pages 236–243. Morgan Kaufmann, San Mateo, CA.

Vestergaard, O., and Sand-Jensen, K. (2000): Aquatic macrophyte richness in Danish lakes in relation to alkalinity, transparency, and lake area. *Canadian Journal of Fisheries and Aquatic Sciences*, 57(10): 2022–2031.