Applied Intelligence 13, 7–17, 2000

Predicting Chemical Parameters of River Water Quality from Bioindicator Data

SAŠO DŽEROSKI AND DAMJAN DEMŠAR
Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia

JASNA GRBOVIĆ
Hydrometeorological Institute, Vojkova 1b, SI-1000 Ljubljana, Slovenia

Abstract. We address the problem of inferring chemical parameters of river water quality from biological ones. This task is important for enabling selective chemical monitoring of river water quality. We apply machine learning, in particular regression tree induction, to biological and chemical data on the water quality of Slovenian rivers. Regression trees are constructed that predict values of chemical parameters from data on the presence of bioindicator taxa at the species and family levels.

Keywords: bioindicators, machine learning, regression trees, rivers, water quality

1. Introduction

The quality of surface waters, including rivers, depends on their physical, chemical and biological properties. The latter are reflected by the types of living organisms present in the water and their density (this includes the structure of the community and its diversity). Based on the above properties, surface waters are classified into (one of) several quality classes which indicate the suitability of the water for different kinds of use.

Since Kolkwitz and Marsson [1] first proposed the use of biota as a means of monitoring the quality of natural waters, many different methods for mapping biological data to discrete quality classes or continuous scales have been developed [2]. Most of these approaches use indicator organisms (bioindicators), which have well-known ecological requirements and are selected for their sensitivity/tolerance to various kinds of pollution. Given a biological sample, information on the presence and density of all indicator organisms present in the sample is usually combined to derive a biological index that reflects the quality of the water at the site where the sample was taken.

Bioindicators can be identified at different taxonomical levels, e.g., at the species level or the family level. A family, a species, or any other taxonomical group can be referred to as a taxon (plural: taxa). In the Saprobic System [1], bioindicators are identified at the species level, which is more demanding in terms of sample processing effort, but also gives a more precise picture of the water quality. The Saprobic Index, based on the Saprobic System, is used to assess water quality in Germany, Austria and Slovenia, among other countries. Family-level identification is carried out for calculating the Biological Monitoring Working Party Score [3], abbreviated as BMWP, and its derivative Average Score Per Taxon (ASPT), which are used in the United Kingdom.

It is well known that the physical and chemical properties give a limited picture of water quality at a particular point in time, while the biota (living organisms) act as continuous monitors of water quality over a period of time [4]. This has increased the relative importance of biological methods for monitoring water quality [2].

The relation between biological and chemical parameters of river water quality is an important and largely open research topic. We have already applied machine learning to the task of inferring biological parameters from chemical ones by learning rules that predict the presence of individual bioindicator taxa from the values of chemical measurements [5, 6]. In this paper, we address the task of inferring chemical

parameters from biological ones. In particular, we learn to predict values of individual chemical parameters from data on the presence of bioindicator taxa. We first use bioindicator data at the species level, then bioindicator data at the family level obtained by aggregating the data at the species level. The machine learning approach of regression tree induction is used, since the chemical measurements are real-valued.

The problem of inferring chemical parameters from biological ones is practically relevant, especially in countries where extensive biological monitoring is conducted. Regular monitoring for a very wide range of chemical pollutants would be very expensive, if not impossible. On the other hand, biological samples may, for example, reflect an increase in pollution and indicate likely causes or sources of (chemical) pollution.

The remainder of the paper is organized as follows. Section 2 describes the measured biological and chemical data on the water quality of Slovenian rivers, as well as the experimental setup. The results of the experiments in learning regression trees with these data are presented in Section 2.3. Section 3 describes how data on bioindicator presence at the species level are aggregated to obtain data at the family level and describes the experiments in learning regression trees with family level data. Section 4 compares the predictive power of regression trees with that of more traditional prediction methods, namely linear regression and nearest neighbor methods. Section 5 concludes with a discussion and some directions for further work.

2. Experiments with Species Level Data

2.1. The Data

The data about Slovenian rivers 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 the six-year period from 1990 to 1995. Biological samples are taken twice a year, in summer and 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 sixteen 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, oxygen saturation, 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 frequency of occurrence (density) of each present taxon is recorded by an expert biologist at three different qualitative levels, where 1 means the taxon occurs incidentally, 3-frequently, and 5-abundantly. The taxa are identified mostly at the species level, so a sample might state that *Tubifex tubifex* was present at abundance level 3. Sometimes, however, taxa might be identified to the genus level only (*Tubifex sp.*) or even the family level only (*Tubificidae*). In total, 1061 water samples were available on which both physical/chemical and biological analyses were performed: all of the experiments presented here were conducted using these samples.

2.2. The Experiments

Approximately 850 different taxa appear in the biological samples. The 415 taxa that appear in at least ten samples were used as attributes (independent variables), while each of the sixteen physical and chemical parameters was used as a class (dependent variable). In this way, sixteen different learning problems were formulated.

As the physical and chemical parameters are real-valued, we used the system M5.1 [7] to induce regression trees for each of the sixteen problems. Ordinary regression trees (with constant predictions in the leaves) and model trees (with linear models in the leaves) were considered.

The default parameters of M5.1 were used in all experiments. For each problem, the following methodology was employed. First, construct a single regression tree from the entire dataset. This tree is shown to domain experts to verify that it contains useful domain knowledge.

An example regression tree induced from the whole dataset is given in Table 1. It predicts the chemical oxygen demand ($K_2Cr_2O_7$) from species-level bioindicator data.

To estimate the performance of the trees induced from the whole dataset on unseen data, ten-fold cross-validation is conducted. The performance, in terms of correlation between the actual values and the predictions, is averaged over the ten-folds. These average results are given in Table 2.

Table 1. A regression tree for predicting chemical oxygen demand ($K_2Cr_2O_7$) induced from species level data.

```
OLIGOCHAETA Lumbriculus variegatus <= 1 :
|   BACTERIA Sphaerotilus natans <= 3 :
|   |   DIPTERA Chironomus thummi > 1 : AV 29.64 (19)
|   |   DIPTERA Chironomus thummi <= 1 :
|   |   |   BACILLARIOPHYTA Nitzschia palea <= 0 :
|   |   |   |   OLIGOCHAETA Tubifex sp. > 1 : AV 11.72 (39)
|   |   |   |   OLIGOCHAETA Tubifex sp. <= 1 :
|   |   |   |   |   COLEOPTERA Elmis sp. > 0 : AV 4.49 (197)
|   |   |   |   |   COLEOPTERA Elmis sp. <= 0 :
|   |   |   |   |   |   HIRUDINEA Erpobdella octoculata > 0 : AV 10.21 (35)
|   |   |   |   |   |   HIRUDINEA Erpobdella octoculata <= 0 :
|   |   |   |   |   |   |   BACILLARIOPHYTA Diatoma hiemale v.mesodon > 1 : AV 2.86 (25)
|   |   |   |   |   |   |   BACILLARIOPHYTA Diatoma hiemale v.mesodon <= 1 :
|   |   |   |   |   |   |   |   BACTERIA Sphaerotilus natans > 0 : AV 9.40 (26)
|   |   |   |   |   |   |   |   BACTERIA Sphaerotilus natans <= 0 :
|   |   |   |   |   |   |   |   |   AMPHIPODA Gammarus fossarum <= 0 : AV 4.31 (68)
|   |   |   |   |   |   |   |   |   AMPHIPODA Gammarus fossarum > 0 : AV 6.91 (57)
|   |   |   BACILLARIOPHYTA Nitzschia palea > 0 :
|   |   |   |   PLECOPTERA Leuctra sp. <= 0 :
|   |   |   |   |   BACILLARIOPHYTA Nitzschia sigmoidea > 0 : AV 8.96 (84)
|   |   |   |   |   BACILLARIOPHYTA Nitzschia sigmoidea <= 0 :
|   |   |   |   |   |   BACTERIA Sphaerotilus natans <= 0 : AV 11.17 (125)
|   |   |   |   |   |   BACTERIA Sphaerotilus natans > 0 :
|   |   |   |   |   |   |   PISCES > 0 : AV 22.18 (20)
|   |   |   |   |   |   |   PISCES <= 0 :
|   |   |   |   |   |   |   |   BACILLARIOPHYTA Nitzschia palea <= 3 : AV 12.69 (96)
|   |   |   |   |   |   |   |   BACILLARIOPHYTA Nitzschia palea > 3 : AV 21.63 (17)
|   |   |   |   PLECOPTERA Leuctra sp. > 0 :
|   |   |   |   |   BACTERIA Sphaerotilus natans <= 1 : AV 7.05 (133)
|   |   |   |   |   BACTERIA Sphaerotilus natans > 1 : AV 13.21 (19)
|   BACTERIA Sphaerotilus natans > 3 :
|   |   AMPHIPODA Gammarus fossarum <= 0 : AV 38.24 (50)          <----------**2**---------->
|   |   AMPHIPODA Gammarus fossarum > 0 :
|   |   |   TRICHOPTERA Rhyacophila sp. <= 0 : AV 26.25 (26)
|   |   |   TRICHOPTERA Rhyacophila sp. > 0 : AV 10.02 (10)
OLIGOCHAETA Lumbriculus variegatus > 1 :
|   BACILLARIOPHYTA Navicula sp. <= 0 : AV 111.47 (6) <----------**1**---------->
|   BACILLARIOPHYTA Navicula sp. > 0 : AV 8.60 (9)
```

2.3. *Results*

The correlation r between the values of parameters predicted by the induced trees and the actual parameter values is given in Table 2: the second column lists the correlation for model trees with linear models in the leaves, the third for ordinary regression trees with constant predictions in the leaves. There is only a slight difference between the two (in favor of the former) and given that trees with linear models in the leaves are more difficult to interpret we only consider the ordinary regression tree results in the following.

Table 2. Correlation between actual chemical parameter values and values predicted by regression/model trees induced from species level data.

Parameter predicted	Model trees r	Regression tree r
BOD	0.659	0.652
Cl	0.590	0.570
CO_2	0.426	0.405
Electrical conductivity	0.581	0.539
$K_2Cr_2O_7$	0.624	0.602
$KMnO_4$	0.558	0.542
NH_4	0.647	0.664
NO_2	0.387	0.373
NO_3	0.406	0.352
O_2	0.542	0.484
Alkalinity (pH)	0.409	0.397
PO_4	0.457	0.461
Oxygen saturation	0.494	0.424
SiO_2	0.487	0.411
Water temperature	0.597	0.561
Total hardness	0.560	0.475

Highest correlations (above 0.6) are achieved when predicting chemical oxygen demand, ammonia and biological oxygen demand. The corresponding trees induced from the entire dataset are given in Tables 1, 3 and 4, respectively.

Highest chemical oxygen demand ($K_2Cr_2O_7$) of (111.47 mg/l) is predicted if the taxon *Lumbriculus variegatus* occurs frequently or abundantly and the taxon *Navicula sp.* does not occur in the sample (leaf **1** in Table 1).

A high value (38.24 mg/l) is also predicted if *Lumbriculus variegatus* occurs at most incidentally, *Sphaerotilus natans* is abundant, and *Gammarus fossarum* does not occur (**2**).

Highest ammonia concentrations are predicted when *Chironomus thummi* occurs frequently or abundantly: 13.35 mg/l if *Sphaerotilus natans* does not occur in the sample (leaf **3** in Table 3) or 6.72 mg/l if it does and *Beggiatoa alba* occurs frequently or abundantly (**4**). Very high ammonia concentration (5.94 mg/l) is also predicted if *Chironomus thummi* does not occur or occurs incidentally, but *Sphaerotilus natans* occurs abundantly, while *Diatoma vulgare* and *Cymbella ventricosa* do not occur in the sample (**5**).

Finally, high BOD is predicted if *Sphaerotilus natans* is frequent or abundant (subtree below **6** in Table 4). BOD is especially high (26.46 mg/l) if in addition *Gammarus fossarum* is absent and *Nitzschia palea* occurs at most incidentally (leaf **7**). The highest value for BOD (27.91 mg/l) is predicted when *Sphaerotilus natans* is not abundant, *Chironomus thummi* is frequent or abundant, and *Cyclotella meneghiniana* is present in the sample (leaf **8** in Table 4).

In general, the trees are in agreement with expert knowledge. The taxa *Lumbriculus variegatus*, *Chironomus thummi*, and *Sphaerotilus natans* emerge as the strongest indicators of heavily polluted waters (they appear at the top of the regression trees). Their presence (and especially abundance) indicates very high chemical oxygen demand, ammonia concentration, and BOD, respectively. The taxa *Beggiatoa alba* and *Cyclotella meneghiniana* are also used as indicators of heavily polluted waters. The taxon *Gammarus fossarum* is used as an indicator of clean to mildly polluted waters. Finally, the taxa *Navicula sp.*, *Cymbella ventricosa*, *Diatoma vulgare*, and *Nitzschia palea* are used as indicators of moderately polluted to polluted (but not heavily polluted) waters.

Some expectations of the domain experts were, however, not fulfilled. For example, caddis flies do not appear in the ammonia tree, despite their tolerance of high ammonia concentrations. Also, *Tubifex tubifex* was expected to play a key role in the BOD tree. Note, however, that the taxon *Lumbriculus variegatus* of the same class (OLIGOCHAETA) plays such a role and that the taxon *Tubifex sp.* also appears in the tree. It is possible that not enough cases of *Tubifex tubifex* were identified to the species level.

3. Experiments with Family Level Data

While species level bioindicator data are collected for the Saprobic System [1], bioindicators are only identified to family level for the BMWP Score [3]. In this section, we examine the effect of using family level instead of species level data on the performance of regression trees when predicting chemical parameters. To this end, we aggregated data on bioindicator presence at the species level to obtain data on bioindicator data at the family level.

3.1. The Data

Taxonomic domain knowledge specifying that certain species belong to certain families was provided by a riverine biology expert (J. Grbović). For example, the taxa *Tubifex tubifex*, *Tubifex sp.*, *Limnodrilus*

Table 3. A regression tree for predicting the NH_4 concentration induced from species level data.

```
DIPTERA Chironomus thummi <= 1 :
|   BACTERIA Sphaerotilus natans <= 3 :
|   |   HIRUDINEA Helobdella stagnalis <= 1 :
|   |   |   PLECOPTERA Leuctra sp. > 0 : AV 0.19 (350)
|   |   |   PLECOPTERA Leuctra sp. <= 0 :
|   |   |   |   CYANOPHYTA Oscillatoria sp. <= 1 :
|   |   |   |   |   OLIGOCHAETA Tubifex tubifex > 0 : AV 1.10 (17)
|   |   |   |   |   OLIGOCHAETA Tubifex tubifex <= 0 :
|   |   |   |   |   |   ISOPODA Asellus aquaticus > 1 : AV 0.57 (56)
|   |   |   |   |   |   ISOPODA Asellus aquaticus <= 1 :
|   |   |   |   |   |   |   OLIGOCHAETA Tubifex sp. > 3 : AV 0.84 (19)
|   |   |   |   |   |   |   OLIGOCHAETA Tubifex sp. <= 3 :
|   |   |   |   |   |   |   |   COLEOPTERA Elmis sp. <= 0 : AV 0.31 (316)
|   |   |   |   |   |   |   |   COLEOPTERA Elmis sp. > 0 : AV 0.18 (147)
|   |   |   |   CYANOPHYTA Oscillatoria sp. > 1 :
|   |   |   |   |   CHLOROPHYTA Stigeoclonium tenue <= 1 : AV 0.53 (22)
|   |   |   |   |   CHLOROPHYTA Stigeoclonium tenue > 1 : AV 2.85 (5)
|   |   HIRUDINEA Helobdella stagnalis > 1 :
|   |   |   DIPTERA Simulium sp. <= 0 : AV 3.59 (9)
|   |   |   DIPTERA Simulium sp. > 0 : AV 0.29 (10)
|   BACTERIA Sphaerotilus natans > 3 :
|   |   BACILLARIOPHYTA Diatoma vulgare > 0 : AV 0.82 (50)
|   |   BACILLARIOPHYTA Diatoma vulgare <= 0 :
|   |   |   BACILLARIOPHYTA Cymbella ventricosa <= 0 : AV 5.94 (13) <----------**5**---------->
|   |   |   BACILLARIOPHYTA Cymbella ventricosa > 0 : AV 1.46 (8)
DIPTERA Chironomus thummi > 1 :
|   BACTERIA Sphaerotilus natans <= 0 : AV 13.35 (9) <----------**3**---------->
|   BACTERIA Sphaerotilus natans > 0 :
|   |   BACTERIA Beggiatoa alba <= 1 : AV 2.55 (24)
|   |   BACTERIA Beggiatoa alba > 1 : AV 6.72 (6) <----------**4**---------->
```

hoffmeisteri and *Limnodrilus sp.* belong to the family *Tubificidae*. This knowledge was used to aggregate the species level data appropriately. Only benthic macroinvertebrates were considered, corresponding to the British system of biological monitoring, where benthic macroinvertebrates identified to family level are used.

Aggregation is performed as follows. Recall that for the taxa present in a sample the abundance level is recorded. The taxa recorded in the samples are typically below or at the family level (mostly at species, occasionally at genus, exceptionally at family level). For each family, we look at all taxa belonging to the family (according to the expert knowledge) present in a sample and take the maximum abundance level among them. This is then the abundance level of the family for the given sample.

This alleviates a potential weakness of using the original data which is a mix of family, genus, and species level data: in the original data, presence at the family level is recorded only where identification was not carried out down to species level.

3.2. *The Experiments and Results*

The attributes in the second series of experiments were the abundance levels of 137 families (of benthic macroinvertebrates) obtained by aggregating

Table 4. A regression tree for predicting biological oxygen demand (BOD) induced from species level data.

```
BACTERIA Sphaerotilus natans <= 3 :
|   DIPTERA Chironomus thummi <= 1 :
|   |   TRICHOPTERA Rhyacophila sp. <= 0 :
|   |   |   AMPHIPODA Gammarus fossarum <= 0 :
|   |   |   |   CHLOROPHYTA Stigeoclonium tenue > 1 : AV 8.88 (18)
|   |   |   |   CHLOROPHYTA Stigeoclonium tenue <= 1 :
|   |   |   |   |   OLIGOCHAETA Tubifex sp. <= 1 :
|   |   |   |   |   |   PLECOPTERA Leuctra sp. > 0 : AV 1.78 (55)
|   |   |   |   |   |   PLECOPTERA Leuctra sp. <= 0 :
|   |   |   |   |   |   |   BACILLARIOPHYTA Diatoma vulgare > 1 : AV 2.36 (42)
|   |   |   |   |   |   |   BACILLARIOPHYTA Diatoma vulgare <= 1 :
|   |   |   |   |   |   |   |   BACTERIA Sphaerotilus natans > 0 : AV 5.61 (29)
|   |   |   |   |   |   |   |   BACTERIA Sphaerotilus natans <= 0 :
|   |   |   |   |   |   |   |   |   DIPTERA Chironomidae green <= 1 : AV 2.09 (40)
|   |   |   |   |   |   |   |   |   DIPTERA Chironomidae green > 1 : AV 4.13 (31)
|   |   |   |   |   OLIGOCHAETA Tubifex sp. > 1 :
|   |   |   |   |   |   BACILLARIOPHYTA Navicula sp. <= 0 : AV 7.52 (21)
|   |   |   |   |   |   BACILLARIOPHYTA Navicula sp. > 0 : AV 3.08 (15)
|   |   |   AMPHIPODA Gammarus fossarum > 0 :
|   |   |   |   GASTROPODA Sadleriana fluminensis > 0 : AV 1.13 (43)
|   |   |   |   GASTROPODA Sadleriana fluminensis <= 0 :
|   |   |   |   |   CHLOROPHYTA Scenedesmus acutus > 1 : AV 6.00 (5)
|   |   |   |   |   CHLOROPHYTA Scenedesmus acutus <= 1 :
|   |   |   |   |   |   BACTERIA Sphaerotilus natans <= 0 :
|   |   |   |   |   |   |   TRICHOPTERA Limnephilidae > 0 : AV 1.64 (41)
|   |   |   |   |   |   |   TRICHOPTERA Limnephilidae <= 0 :
|   |   |   |   |   |   |   |   CHLOROPHYTA Scenedesmus quadricauda <= 0 : AV 2.42 (111)
|   |   |   |   |   |   |   |   CHLOROPHYTA Scenedesmus quadricauda > 0 : AV 3.43 (50)
|   |   |   |   |   |   BACTERIA Sphaerotilus natans > 0 :
|   |   |   |   |   |   |   BACILLARIOPHYTA Nitzschia sigmoidea <= 0 : AV 4.02 (99)
|   |   |   |   |   |   |   BACILLARIOPHYTA Nitzschia sigmoidea > 0 : AV 2.32 (28)
|   |   TRICHOPTERA Rhyacophila sp. > 0 :
|   |   |   CYANOPHYTA Dactylococcopsis rhaphidioides > 0 : AV 8.41 (8)
|   |   |   CYANOPHYTA Dactylococcopsis rhaphidioides <= 0 :
|   |   |   |   BACILLARIOPHYTA Surirella ovata <= 0 : AV 1.54 (246)
|   |   |   |   BACILLARIOPHYTA Surirella ovata > 0 : AV 2.43 (69)
|   DIPTERA Chironomus thummi > 1 :
|   |   BACILLARIOPHYTA Cyclotella meneghiniana <= 0 : AV 7.47 (12)
|   |   BACILLARIOPHYTA Cyclotella meneghiniana > 0 : AV 27.91 (8) <----------**8**---------->
BACTERIA Sphaerotilus natans > 3 : <----------**6**---------->
|   AMPHIPODA Gammarus fossarum > 0 : AV 9.09 (38)
|   AMPHIPODA Gammarus fossarum <= 0 :
|   |   BACILLARIOPHYTA Nitzschia palea <= 1 : AV 26.46 (14) <----------**7**---------->
|   |   BACILLARIOPHYTA Nitzschia palea > 1 :
|   |   |   BACILLARIOPHYTA Navicula cryptocephala v.cryptoceph. <= 1 : AV 12.55 (23)
|   |   |   BACILLARIOPHYTA Navicula cryptocephala v.cryptoceph. > 1 : AV 20.91 (15)
```

Table 5. Correlations between actual chemical parameter values and values predicted by regression trees induced induced from species and family level data.

	Species level r	Family level r
BOD	0.652	0.424
Cl	0.570	0.444
CO_2	0.405	0.431
Electrical conductivity	0.539	0.447
$K_2Cr_2O_7$	0.602	0.416
$KMnO_4$	0.546	0.340
NH_4	0.664	0.351
NO_2	0.373	0.214
NO_3	0.352	0.279
O_2	0.484	0.332
Alkalinity (pH)	0.397	0.281
PO_4	0.461	0.280
Oxygen saturation	0.424	0.302
SiO_2	0.411	0.359
Water temperature	0.561	0.261
Water hardness	0.475	0.477

the original data as described above. Each of the sixteen physical and chemical parameters was used as a class (dependent variable), thus yielding sixteen learning problems analogous to the ones described in Section 2.2. Only regression trees (no linear models in the leaves) were induced. Otherwise, the same experimental setup was used as for species level data (see Section 2.2).

The correlation r between the values of parameters predicted by the induced trees and the actual parameter values is listed in the third column of Table 5: the second column lists the correlation for trees derived from species level data for comparison.

Correlations for trees using family level data are lower for all chemical parameters, except CO_2. This is understandable, as species level data contain more information on the quality of water. Namely, families comprise a number of species which can have different tolerances to pollutants.

All correlations are below 0.5. Of the three chemical parameters considered in the previous section, the highest drop is observed for ammonia (0.3 difference in correlation). For comparison with Tables 1, 3 and 4, the trees for $K_2Cr_2O_7$, ammonia and BOD induced from the entire dataset are given in Tables 6–8, respectively.

Highest chemical oxygen demand (95.66 mg/l) is predicted if the family *Lumbriculidae* from the class OLIGOCHAETA occurs abundantly (leaf **9** in Table 6). The tree is similar to the one in Table 1, where the abundance of the species *Lumbriculus variegatus* is the most important attribute. The presence of *Tubificidae* indicates higher oxygen demand (subtree below **10**), as does the absence of *Elmidae* (**11**). This is consistent with the fact that the former are indicators of dirty and the latter of clean water.

The tree for predicting the ammonia concentration (Table 7) is substantially different from the tree in Table 3. While the taxon *Chironomus thummi* is apparently quite tolerant of high ammonia concentrations, the *Chironomidae* family comprises many different species with different tolerances to pollution. A recent reappraisal of BMWP scores [8] shows that *Chironomidae* are not a good indicator of water quality for this very reason.

Highest ammonia concentration (9.31 mg/l, leaf **12** in Table 7) is predicted when the family *Baetidae* is absent and the family *Culicidae* from the class of true flies (DIPTERA) is present. While *Baetidae* are considered indicators of relatively clean waters, *Culicidae* are currently not used as bioindicators within the Saprobic System (in Slovenia). The tree suggests that *Culicidae* are tolerant of high ammonia concentrations and indicate low water quality.

In general, the trees are in agreement with expert knowledge. The family *Tubificidae* appears at the root of the tree (is the most important) for predicting BOD from family level data (Table 8), thus directly addressing the expert comments on the species level tree for BOD: *Tubifex tubifex* was namely expected to play a key role in the species level tree, but did not. The bacterium *Sphaerotilus natans* appeared at the top of the species level tree, meaning that it is more indicative of high BOD than *Tubificidae*. Among benthic macroinvertebrates, however, *Tubificidae* are the most indicative of high BOD.

It is worth noting that the trees actually complement the expert knowledge, adding new pieces to the mosaic of water quality. An example of this is the suggestion that *Culicidae* are indicators of low water quality.

4. Comparison with Linear Regression and Nearest Neighbor Prediction

One might argue that using more traditional prediction methods, such as linear regression or nearest neighbor methods is preferrable to the method of regression

Table 6. A regression tree for predicting chemical oxygen demand ($K_2Cr_2O_7$) induced from family level data.

```
familia_OLIGOCHAETA_Lumbriculidae > 3 : AV 95.66 (7) <----------**9**---------->
familia_OLIGOCHAETA_Lumbriculidae <= 3 :
|   familia_OLIGOCHAETA_Tubificidae <= 1 :
|   |   familia_COLEOPTERA_Elmidae <= 0 : <----------**11**---------->
|   |   |   familia_EPHEMEROPTERA_Heptageniidae <= 0 :
|   |   |   |   familia_HIRUDINEA_Erpobdellidae <= 0 : AV 10.46 (133)
|   |   |   |   familia_HIRUDINEA_Erpobdellidae > 0 : AV 14.88 (113)
|   |   |   familia_EPHEMEROPTERA_Heptageniidae > 0 :
|   |   |   |   familia_PLECOPTERA_Nemouridae > 1 : AV 3.69 (28)
|   |   |   |   familia_PLECOPTERA_Nemouridae <= 1 :
|   |   |   |   |   familia_TRICHOPTERA_Hydropsychidae <= 0 : AV 6.51 (45)
|   |   |   |   |   familia_TRICHOPTERA_Hydropsychidae > 0 : AV 10.52 (63)
|   |   familia_COLEOPTERA_Elmidae > 0 :
|   |   |   familia_ISOPODA_Asellidae > 0 : AV 11.22 (52)
|   |   |   familia_ISOPODA_Asellidae <= 0 :
|   |   |   |   familia_HIRUDINEA_Erpobdellidae <= 0 : AV 4.89 (307)
|   |   |   |   familia_HIRUDINEA_Erpobdellidae > 0 : AV 7.53 (66)
|   familia_OLIGOCHAETA_Tubificidae > 1 : <----------**10**---------->
|   |   familia_AMPHIPODA_Gammaridae > 0 : AV 13.26 (136)
|   |   familia_AMPHIPODA_Gammaridae <= 0 :
|   |   |   familia_HIRUDINEA_Glossiphoniidae > 0 : AV 17.29 (31)
|   |   |   familia_HIRUDINEA_Glossiphoniidae <= 0 :
|   |   |   |   familia_OLIGOCHAETA_Lumbricidae > 0 : AV 15.32 (14)
|   |   |   |   familia_OLIGOCHAETA_Lumbricidae <= 0 :
|   |   |   |   |   familia_DIPTERA_Chironomidae <= 1 : AV 18.24 (14)
|   |   |   |   |   familia_DIPTERA_Chironomidae > 1 : AV 36.21 (52)
```

trees that we proposed to use. We briefly present here a comparison of the three approaches on the problem of predicting chemical parameters from species level bioindicator data.

The M5.1 program actually also includes facilities for linear regression and nearest neighbor prediction. We used these capabilities of M5.1 to perform the experiments reported here. Standard linear regression and 3-NN prediction (where the three nearest neighbors of an instance are used to predict the value of that instance) are implemented in M5.1. The correlation between actual and predicted parameter values for each of the three approaches is given in Table 9. The same experimental setup (ten-fold cross-validation) was used for all three approaches.

We see that the three approaches are very close in terms of predictive performance, especially regression trees and nearest neighbor prediction. The regression trees, however, have the advantage of producing reasonably-sized structured generalizations of the input data that are understandable as well, as demonstrated by the expert comments on the trees in Sections 2 and 3. It is easier to interpret a regression tree of a moderate size (as in the Tables in this paper) than 415 coefficients in a linear equation. The nearest neighbor method, on the other hand, produces no generalization of the input data at all.

5. Discussion

This paper addresses the problem of inferring chemical parameters of river water quality from biological ones by using regression tree induction. Initial experiments indicate that ammonia concentration, biological oxygen demand and chemical oxygen demand can be predicted relatively successfully from bioindicator data. One should bear in mind that changes in the biota may be caused by short-term fluctuations of chemical

Table 7. A regression tree for predicting the NH_4 concentration induced from family level data.

```
familia_EPHEMEROPTERA_Baetidae <= 0 :
|   familia_DIPTERA_Culicidae > 0 : AV 9.31 (10) <----------**12**---------->
|   familia_DIPTERA_Culicidae <= 0 :
|   |   familia_AMPHIPODA_Gammaridae <= 0 :
|   |   |   familia_DIPTERA_Chironomidae <= 1 : AV 0.79 (46)
|   |   |   familia_DIPTERA_Chironomidae > 1 :
|   |   |   |   familia_OLIGOCHAETA_Tubificidae <= 1 :
|   |   |   |   |   familia_HIRUDINEA_Glossiphoniidae <= 0 : AV 0.32 (31)
|   |   |   |   |   familia_HIRUDINEA_Glossiphoniidae > 0 : AV 3.38 (9)
|   |   |   |   familia_OLIGOCHAETA_Tubificidae > 1 :
|   |   |   |   |   familia_OLIGOCHAETA_Naididae <= 0 : AV 4.32 (37)
|   |   |   |   |   familia_OLIGOCHAETA_Naididae > 0 : AV 1.37 (13)
|   |   familia_AMPHIPODA_Gammaridae > 0 :
|   |   |   familia_DIPTERA_Orthocladiinae <= 1 : AV 0.32 (111)
|   |   |   familia_DIPTERA_Orthocladiinae > 1 : AV 1.30 (8)
familia_EPHEMEROPTERA_Baetidae > 0 :
|   familia_OLIGOCHAETA_Tubificidae <= 3 :
|   |   familia_TRICHOPTERA_Rhyacophilidae > 0 : AV 0.18 (326)
|   |   familia_TRICHOPTERA_Rhyacophilidae <= 0 :
|   |   |   familia_AMPHIPODA_Gammaridae <= 0 :
|   |   |   |   familia_COLEOPTERA_Elmidae > 0 : AV 0.23 (54)
|   |   |   |   familia_COLEOPTERA_Elmidae <= 0 :
|   |   |   |   |   familia_EPHEMEROPTERA_Baetidae <= 1 : AV 1.04 (38)
|   |   |   |   |   familia_EPHEMEROPTERA_Baetidae > 1 : AV 0.41 (44)
|   |   |   familia_AMPHIPODA_Gammaridae > 0 :
|   |   |   |   familia_ISOPODA_Asellidae <= 0 : AV 0.24 (225)
|   |   |   |   familia_ISOPODA_Asellidae > 0 : AV 0.52 (72)
|   familia_OLIGOCHAETA_Tubificidae > 3 :
|   |   familia_AMPHIPODA_Gammaridae <= 0 : AV 2.29 (21)
|   |   familia_AMPHIPODA_Gammaridae > 0 : AV 0.33 (16)
```

parameter values, meaning that it is impossible to completely determine the latter from the former. Nevertheless, our work is a step towards enabling selective chemical monitoring of river water quality.

We show that species level biological data carry much more information on the chemical water quality than family level data. This is especially true for families that include species with large differences in tolerance to pollutants, such as *Chironomidae*. The largest drop in predictive performance when moving from species to family level data was observed for ammonia concentrations.

While one might expect that combining species level data with properly aggregated family level data could further improve the predictive performane of regression trees on the problem at hand, it turns out that it doesn't. We have conducted experiments in the same general setup described above where both the species and the family level data were used, but this resulted in no performance gains over the species level data results. This is in a way understandable, since all the information on the biological state of the water is already in the species level data.

It should be noted that the biological state of a river at a given time is influenced by the chemical state over some period of time up to the given point. Further work should thus take into account several chemical measurements preceding a given biological sample.

Table 8. A regression tree for predicting the biological oxygen demand (BOD) induced from family level data.

```
familia_OLIGOCHAETA_Tubificidae <= 1 :
|   familia_COLEOPTERA_Elmidae <= 0 :
|   |   familia_EPHEMEROPTERA_Heptageniidae <= 0 : AV 4.38 (246)
|   |   familia_EPHEMEROPTERA_Heptageniidae > 0 :
|   |   |   familia_PLECOPTERA_Nemouridae <= 1 : AV 2.63 (108)
|   |   |   familia_PLECOPTERA_Nemouridae > 1 : AV 1.30 (28)
|   familia_COLEOPTERA_Elmidae > 0 :
|   |   familia_GASTROPODA_Hydrobiidae > 0 : AV 1.15 (109)
|   |   familia_GASTROPODA_Hydrobiidae <= 0 :
|   |   |   familia_PLECOPTERA_Leuctridae <= 0 : AV 2.61 (154)
|   |   |   familia_PLECOPTERA_Leuctridae > 0 : AV 1.67 (163)
familia_OLIGOCHAETA_Tubificidae > 1 :
|   familia_AMPHIPODA_Gammaridae > 0 : AV 4.05 (137)
|   familia_AMPHIPODA_Gammaridae <= 0 :
|   |   familia_DIPTERA_Psychodidae > 0 : AV 37.24 (5)
|   |   familia_DIPTERA_Psychodidae <= 0 :
|   |   |   familia_HIRUDINEA_Glossiphoniidae <= 0 : AV 12.89 (80)
|   |   |   familia_HIRUDINEA_Glossiphoniidae > 0 : AV 7.02 (31)
```

Table 9. Correlations between actual chemical parameter values and values predicted by regression trees, linear equations and nearest neighbor induced from species level data.

	RT	LR	NN
BOD	0.652	0.592	0.577
Cl	0.570	0.618	0.595
CO_2	0.405	0.409	0.398
Electrical conductivity	0.539	0.571	0.553
$K_2Cr_2O_7$	0.602	0.596	0.614
$KMnO_4$	0.546	0.485	0.577
NH_4	0.664	0.487	0.534
NO_2	0.373	0.318	0.448
NO_3	0.352	0.316	0.360
O_2	0.484	0.544	0.551
Alkalinity (pH)	0.397	0.382	0.427
PO_4	0.461	0.352	0.569
Oxygen saturation	0.424	0.454	0.464
SiO_2	0.411	0.509	0.445
Water temperature	0.561	0.588	0.479
Water hardness	0.475	0.472	0.533
Average	0.498	0.481	0.508

RT = regression trees; LR = linear regression; NN = nearest neighbor.

For some chemical parameters, it might turn out that their cumulative effects (average values over the given period) are more pronounced and thus their average values are easier to predict. For others, maximum (e.g., for ammonia) or minimum (e.g., for oxygen) values might be more relevant.

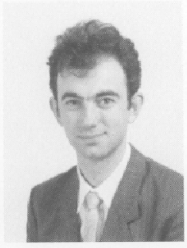
In further work, we will try to predict average, maximum or minimum values of chemical parameters over a period of time preceding the biological sampling, rather than the values of chemical parameters at the single point in time (of biological sampling) as done in this paper. As the abundance of some bioindicators can be affected by the seasons, seasonal effects should also be taken into account in further work.

Acknowledgments

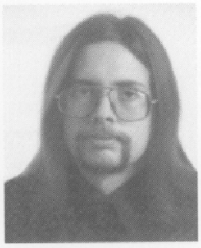
The research described in this paper was supported by the Slovenian Ministry of Science and Technology. The Hydrometeorological Institute of Slovenia provided the biological, physical and chemical data on Slovenian rivers used in this study. Thanks are due to William J. Walley and Herbert A. Hawkes for comments on the regression trees derived from species level data.

References

1. R. Kolkwitz and M. Marsson, "Grundsatze für die biologische Beurteilung des Wassers nach seiner Flora und Fauna," *Mitt. Prüfungsanst. Wasserversorg. Äbwasserein*, vol. 1, pp. 33–72, 1902.
2. N. De Pauw and H.A. Hawkes, "Biological monitoring of river water quality," in *Proc. Freshwater Europe Symposium on River Water Quality Monitoring and Control*, Aston University, Birmingham, 1993, pp. 87–111.
3. ISO-BMWP, "Assessment of the biological quality of rivers by a macroinvertebrate score," ISO/TC147/SC5/WG6/N5, International Standards Organization, 1979.
4. J. Cairns, W.A. Douglas, F. Busey, and M.D. Chaney, "The sequential comparison index—a simplified method for non-biologists to estimate relative differences in biological diversities in stream pollution studies," *J. Wat. Pollut. Control Fed.*, vol. 40, pp. 1607–1613, 1968.
5. S. Džeroski and J. Grbović, "Knowledge discovery in a water quality database," in *Proc. First International Conference on Knowledge Discovery and Data Mining*, AAAI Press: Menlo Park, CA, 1995, pp. 81–86.
6. S. Džeroski, J. Grbović, W.J. Walley, and B. Kompare, "Using machine learning techniques in the construction of models, Part II: Rule induction," *Ecological Modelling*, vol. 95, pp. 95–111, 1997.
7. J.R. Quinlan, "Combining instance-based and model-based learning," in *Proc. Tenth International Conference on Machine Learning*, Morgan Kaufmann: San Mateo, CA, 1993, pp. 236–243.
8. W.J. Walley and H.A. Hawkes, "A computer-based reappraisal of the Biological Monitoring Working Party scores using data from the 1990 river quality survey of England and Wales," *Water Research*, vol. 30, no. 9, pp. 2086–2094, 1996.



Sašo Džeroski is a research associate at the Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia (since 1989). He received his PhD in computer science in 1995 from the Faculty of Electrical Engineering and Computer Science, University of Ljubljana. He has held visiting researcher positions at the Turing Institute, Glasgow (UK), Katholieke Universiteit Leuven (Belgium), German National Research Center for Computer Science (GMD), Sankt Augustin (Germany) and the Foundation for Research and Technology-Hellas (FORTH), Heraklion (Greece). His research interest is in machine learning and knowledge discovery in databases, in particular inductive logic programming and its applications and knowledge discovery in environmental databases.



Damjan Demšar is a reaserch assistant at the Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia. He received his MS in computer science in 1999 from the Faculty of Computer and Information Science, University of Ljubljana, where he is currently a PhD student. His research interest is mainly in the field of machine learning and its applications in ecological domains.

Jasna Grbović is an advisor to the director and head of the Biological Laboratory at the Hydrometeorological Institute of Slovenia, Ljubljana, Slovenia. She received her PhD in biology in 1994 from the Faculty of Biology, University of Ljubljana, Slovenia. She was a research assistant in the field of water ecology at the National Institute of Chemistry, Ljubljana, Slovenia from 1973 to 1992, when she moved to her present position. She has since been a visiting reasercher at the Department of Industrial Chemistry, University of Padova, Italy. Her current research includes biological evaluation of water quality, recognition of main ecological factors other than pollution which affect or limit the distribution of species and/or biocenoses, the study of correlations among physical, chemical, bacteriological and ecological aspects of water quality and establishing criteria for classification of surface running water quality.