Knowledge discovery in a water quality database

Sašo Džeroski
Jožef Stefan Institute
Jamova 39, 61111 Ljubljana, Slovenia
Email: saso.dzeroski@ijs.si

Jasna Grbović
Hydrometeorological Institute of Slovenia
Vojkova 1b, 61000 Ljubljana, Slovenia

Abstract

We apply rule induction to mine for knowledge in a database which stores data obtained by monitoring the water quality of the rivers in Slovenia. The database contains measurement data about the physical and chemical properties of the water at different measurement sites, as well as data about the presence of living organisms. Taken together, the above data reflect the quality of the water: the physical and chemical properties influence the living organisms, which in turn give an overall picture of the water quality over a period of time. We address two problems: (a) analysis of the influence of physical and chemical indicators of water quality on the presence of selected bioindicators, and (b) classification into quality classes based on either bioindicator or physical and chemical indicator data. The learned rules are evaluated by a river biology expert, but also in terms of their performance on unseen cases.

Introduction

The quality of surface waters, including rivers, depends on their physical, chemical and biological properties. The latter are reflected by the diversity and density of living organisms in the water. 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.

According to current legislation on water classification in Slovenia (Official 1978), the water belongs to the first (best) quality class if it is suitable for drinking, bathing and fisheries. Second class water is suitable for fisheries and recreation, including bathing; after simple treatment (coagulation, filtration, disinfection) it can be used for industrial purposes, even in the food industry. Third class waters can be used for irrigation and (after conditioning) in the industry, except the food industry. Water of the fourth (worst) quality class can be used only for purposes less demanding than the above ones and after appropriate treatment.

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. This has increased the importance of biological methods for monitoring water quality (De Pauw & Hawkes 1993). Since Kolkwitz and Marsson (1902), who 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 (De Pauw & Hawkes 1993; Grbović 1994). Unfortunately, the rationales behind these methods are typically ad hoc and highly subjective and their reliabilities have been less than desired (Walley, Boyd, & Hawkes 1992).

Slovenian water authorities use the saprobic index method, as introduced by Pantle and Buck (1955) and modified by Zelinka and Marvan (1961), to map biological data to a continuous scale. The saprobic index derived from a given water sample is thus a single number that reflects the quality of the water. Depending on the value of the saprobic index water can be classified in four basic classes and three intermediate classes, i.e., altogether seven classes: 1., 1.-2., 2., 2.-3., 3., 3.-4., and 4. Class 1. corresponds to clean waters and class 4. to heavily polluted waters. The four basic classes correspond to the legislation defined classes, but are somewhat different, as the latter rely mainly on chemical properties.

The saprobic index is calculated as a weighted average of the densities of a selected set of living organism families (or other taxonomical units, referred to as taxa). The taxa used are such that their biology, importance and ecological role is known. Such taxa are called bioindicators, since they reflect the overall water quality as affected by physical and chemical influences over a period of time. The ecological role and water quality importance is not known for many taxa and may furthermore differ from country to country (Grbović 1994). Little is also known about the influence of physical and chemical water properties on many taxa. From an ecological and water quality point of view, these are important research topics.

The paper describes experiments in mining a water quality database for knowledge on two topics:

(a) the influence of physical and chemical water properties on selected bioindicator taxa and (b) the classification (quality assessment) of water samples, based on either bioindicator data or data on the physical and chemical properties of the sample. The task of water quality classification is often referred to as the task of interpretation of, e.g., biological samples (Walley, Boyd, & Hawkes 1992). The data we used comes from the Hydrometeorological Institute of Slovenia (Hidrometereološki Zavod Republike Slovenije, abbr. 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 four year period, from 1990 to 1993. 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 several different parameters, such as dissolved oxygen and hardness, while the biological samples include a list of all taxa present at the sampling site and their density. The density (abundance level) 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. Biological samples include the corresponding saprobic index value and the corresponding quality class as determined by the index. In total, 698 water samples were available on which both physical/chemical and biological analyses were performed: our experiments were conducted using these samples.

Given the data described above, we used a rule induction system to mine for knowledge. We formulated several learning problems: analysis of the influence of selected physical and chemical water properties on the presence of selected taxa; water quality classification starting from a selected set of bioindicators; and water quality classification based on a a selected set of physical and chemical properties. In the remainder of the paper, we first describe the methodology used to learn and evaluate rules and then present the results for each of the learning problems. The evaluation of induced rules comprises comments by Jasna Grbović, an expert biologist that performs analyses of biological samples at HMZ and has rich knowledge on the ecology of plants and animals found in Slovenian rivers, as well as the classification accuracy and information score of the rules (estimated on unseen cases).

The methodology of rule induction

To induce rules from the given examples, we used an extended version (Džeroski, Cestnik, & Petrovski 1993) of CN2 (Clark & Boswell 1991). The extended version can use the m-estimate (Cestnik 1990) for estimating rule accuracy: the accuracy estimates are used as values of the search heuristic in CN2. The m-estimate gives more reliable probability estimates and allows for different levels of fitting the training examples: smaller values of the parameter m (which is a positive real number) correspond to closer fitting (Cestnik 1990). The rationale behind the parameter m is to allow for better noise-handling. Another feature of the extended version of CN2 is the possibility to measure the information score (Kononenko & Bratko 1991) of induced rules. The information score is a performance measure which is not biased by the prior class distribution. It accounts for the possibility to achieve high accuracy easily in domains with a very likely majority class: classifying into the majority class all the time gives a zero information score.

CN2 was used to induce sets of unordered rules. The rules were required to be highly significant (at the 99% level) and thus reliable. Except for the the significance threshold and the search heuristic settings, described below, the parameter settings of CN2 were the default ones (see Clark & Boswell 1991).

To estimate probabilities for the search heuristic (i.e., rule accuracy) we used the **Laplace estimate** and the **m-estimate**. Fifteen different values of the parameter m were tried (0, 0.01, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024), as suggested by earlier experiments (Cestnik 1990; Džeroski, Cestnik, & Petrovski 1993). For a given set of examples, we thus induced 16 sets of rules and chose the best according to the following lexicographic criterion: (1) information score, (2) accuracy, (3) smaller value of the parameter m. The accuracy and the relative information score are estimated on the training set.

This procedure allows us to choose the right level of fitting: overfitting is prevented by applying the significance threshold. Preliminary experiments showed that as the parameter m increases, the accuracy and information score of the induced rules increase until an optimum is reached; further increasing m causes a decrease in the accuracy and information score (Ličan-Milošević 1994). This behavior is illustrated in Figure 1.

Note that a behavior of this kind is obtained only if we use a high significance threshold. If we don't apply a significance threshold then accuracy and information score fall as m grows: this prevents us from being able to choose an appropriate value of m on the training set. Earlier experiments chose an appropriate value for m on the testing set (Cestnik 1990, Džeroski, Cestnik, & Petrovski 1993), which is a methodological flaw.

For each of the learning problems described below, two sets of experiments were performed. The first set induced rules from all 698 examples, aiming to find as much reliable patterns (and hopefully knowledge) as possible. The rules derived in this way were inspected by the expert biologist and evaluated in the light of existing knowledge on riverine ecology and water quality. The second batch of experiments was aimed at evaluating the performance of induced rules in terms of their accuracy and information score on unseen cases. To this end, we split the entire dataset into a training

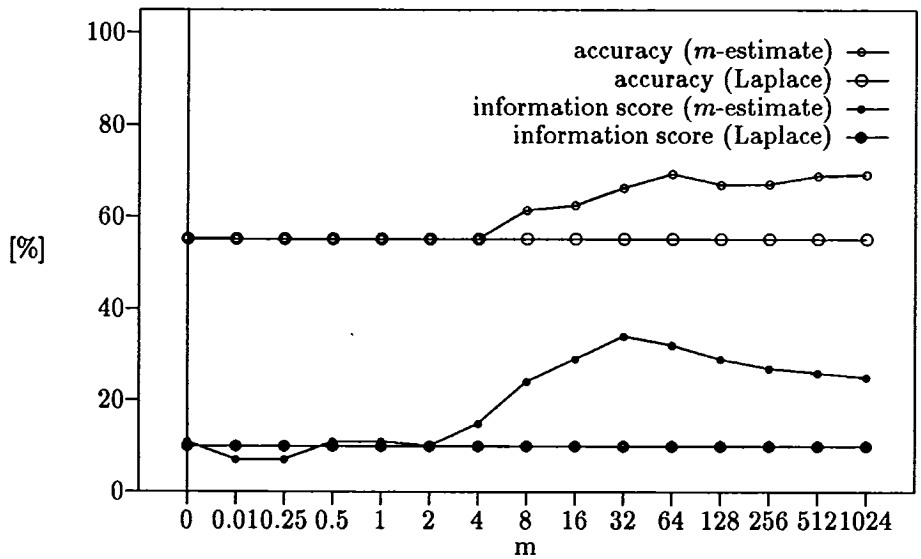

Figure 1: The accuracy and information score of significant rules learned by using different search heuristics.

(70% - 489 cases) and testing (30% - 209 cases) set in ten different ways. For each split, we induced a set of rules according to the above methodology from the training set, then tested the performance of the rules on the test set; the results appearing in the text are the averages over the ten different splits.

The influence of physical and chemical parameters on selected organisms

Plants are more or less influenced by the following physical and chemical parameters (water properties): total hardness, nitrogen compounds (NO_2, NO_3, NH_4), phosphorus compounds (PO_4), silica (SiO_2), iron (Fe), surfactants (detergents), chemical oxygen demand (COD), and biochemical oxygen demand (BOD). The last two parameters indicate the degree of pollution: the first reflects the total amount of degradable matter, while the second reflects the amount of biologically degradable matter. Animals are influenced by a different set of parameters: water temperature, acidity or alkalinity (pH), dissolved oxygen (O_2, saturation of O_2), total hardness, chemical (COD), and biochemical oxygen demand (BOD).

The experiments presented in this section studied the influence of the physical and chemical parameters listed above on ten plant taxa and seven animal taxa. On the plant side, eight kinds of diatomae (BACILLARIOPHYTA: 12200, 13200, 13700, 14200, 17700, 18000, 19400, 21100) and two kinds of green algae (CHLOROPHYTA: 25400, 30400) were studied. The animal taxa chosen for study include worms (OLIGOCHAETA: 37880), crustacea (AMPHIPODA: 49700) and five kinds of insects (50380, 53000, 58900, 59300, 59500). The numbers above are the taxa identification codes from a code-book of taxa that appear in Slovene rivers; this code-book is maintained by HMZ.

For each of the selected taxa we defined an attribute-based learning problem, the attributes being the selected physical and chemical parameters (`Hardness`, `NO2`, `NO3`, `NH4`,`PO4`, `SiO2`, `Fe`, `Detergents`, `COD`, `BOD` for plants; `Temperature`, `PH`, `O2`, `Saturation`, `COD`, `BOD` for animals). The class is the presence of the selected taxon (with values `Present` and `Absent`). Seventeen different learning problems (domains) were thus defined.

We now summarize the experiments with the above learning problems, carried in accordance with the methodology specified above. We first give an overview of the performance of the induced rules, both for rules derived from the whole data set (calculated on the training set) and for rules derived from the ten splits (calculated on the testing set and averaged over the ten splits). We then give excerpts of the expert rule evaluation for selected plant and animal taxa. A detailed description of the selected taxa, the experiments, the performance of induced rules, the selected sets of rules, and the expert evaluation of these rules can be found in the BSc Thesis of Ličan-Milošević (1994).

Table 1 gives an overview of the performance of the rules as evaluated on the whole dataset (W) and the ten splits (P) into a training and testing set. The accuracy on the whole (training) dataset ranges between 66% and 85% (the default accuracy, i.e., the majority class frequency ranges from 50% to 70%), while the information score ranges between 23% and 50%. The rule sets for different taxa comprised 10 to 20 rules, the average rule length was less than five conditions, and a rule covered on average 15 to 45 examples.

Table 1: The performance of rules predicting taxa presence from physical and chemical parameters: D-default accuracy, A-accuracy, I-information score; W-whole dataset, P-average on ten splits.

Plant taxa						
	D		A		I	
Code	W	P	W	P	W	P
12200	55.9	58.4	77.8	65.5	43	20.0
13200	65.8	64.6	77.7	67.7	33	16.6
13700	56.3	59.3	75.6	63.0	38	15.1
14200	59.7	56.5	79.2	64.6	38	14.6
17700	63.2	61.7	78.4	62.6	41	12.1
18000	66.6	71.8	79.8	71.3	41	16.9
19400	60.2	60.8	78.8	69.7	50	29.2
21100	50.0	48.3	69.1	53.1	43	7.0
25400	54.6	59.8	66.3	56.8	34	11.4
30400	69.5	68.4	80.8	67.6	35	5.1
Animal taxa						
	D		A		I	
Code	W	P	W	P	W	P
37880	67.2	66.0	85.1	70.0	49	20.5
49700	56.7	55.0	75.9	64.6	35	17.6
50380	65.9	67.9	77.9	65.7	31	7.0
53000	68.2	68.9	85.1	70.6	47	17.9
58900	55.3	50.2	67.2	53.3	23	1.4
59300	65.2	75.6	75.5	65.4	28	4.6
59500	65.8	64.6	79.8	71.0	46	26.1

While the performance (accuracy) of the rules on the training set is not as high as might be expected, we should bear in mind that the use of a high significance threshold prevents overfitting. More importantly, the physical and chemical parameters at a certain point in time do not determine completely the presence (absence) of a particular taxon: the presence depends on the physical and chemical parameters over a longer period of time, on the life time of the taxon, the water level, and the river bed. To make the problem even harder, some taxa group together very different organisms: an example is the taxon *Chironomidae (green)* (58900), where the lowest information score on the whole dataset was recorded.

The information scores of the rules induced from 70% of the dataset (measured on the remaining 30%) is much lower for all taxa, but remains positive. This means that the rules contain useful information about the influence of physical and chemical parameters on the presence of the taxa. Nevertheless, the accuracy is worse than the default for the taxa 18000, 25400, 30400, 50380, and 59400. In the remainder of the section, we give an excerpt from the expert evaluation of the rules for the diatom *Nitzschia palea (19400)* and the water bug *Elmis sp. (59500)*. The rules for these taxa achieved the highest information scores (29.2% and 26.1%, respectively) on the 70%-30% splits.

The diatom *Nitzschia palea (19400)* is present in 420 of the 698 samples and is the most common species in Slovenian rivers. It is very tolerant to pollution and lives in waters of a wide quality range, from clean to polluted waters. It is characteristic of the water quality class 2.-3. according to the saprobic index and is used as an indicator of polluted waters.

```
% Selected rule predicting the presence
% of the species Nitzschia palea

IF  PO4 > 0.065                IF  NO3 > 1.3
AND Fe < 0.595                 AND NH4 < 0.97
AND COD > 25.5                 AND 13.25 < COD < 16.35
THEN Present [58 0]            THEN Present [36 0]

IF  4.25 < NO3 < 12.35         IF  Hardness > 11.85
AND SiO2 > 1.65                AND NO2 > 0.095
AND Detergents > 0.055         AND NH4 > 0.09
THEN Present [50 0]            THEN Present [82 0]

IF  NO3 < 5.95                 IF  NO2 < 0.005
AND SiO2 > 4.75                AND NO3 < 7.1
AND COD > 7.95                 AND PO4 < 0.125
AND 1.3 < BOD < 42.05          AND Detergents < 0.055
THEN Present [59 0]            AND BOD < 2
                               THEN Absent [0 39]
```

The rules built from the whole dataset confirm that a larger degree of pollution is beneficial to this species. From the 18 rules we list six below, chosen to have large coverage. The numbers in square brackets denote the numbers of examples of each class covered by the rule ([58 0] means that the corresponding rule covers 58 examples of class **Present**, while [0 39] means it covers 39 examples of the class **Absent**). From the rules it is evident that *Nitzschia palea* needs nitrogen compounds, phosphates, silica, and larger amounts of degradable matter (*COD* and *BOD*).

The bugs *COLEOPTERA*, where the taxon *Elmis sp. (59500)* belongs, are quite common on land but rare in water. From the literature and expert experiences it is known that this taxon inhabits clean waters: it is considered an indicator of the quality class 1.-2.

```
% Selected rule predicting the presence
% of the taxon Elmis sp.

IF  O2 < 11.45                 IF  Temperature > 12.75
AND Hardness > 10.35           AND BOD < 0.65
AND COD > 2.15                 THEN Present [8 0]
AND BOD < 1.25
THEN Present [36 0]            IF  23 < COD < 46.45
                               THEN Absent [0 72]
IF  Temperature > 11.75
AND 12.3 < Hardness <14.3      IF  PH > 7.05
AND BOD < 1.75                 AND BOD > 12.15
THEN Present [14 0]            THEN Absent [0 47]
```

From the 17 rules induced, five selected by the expert are listed above. The first rule demands a relatively low quantity of biodegradable matter (pollution) in order for *Elmis sp.* to be present; this has to be even lower as water temperature increases (see the second and the third rule). The last two rules predict that the taxon will be absent if the water is overly polluted as

indicated by the high values of *BOD*, *COD* and *pH*. The rules confirm that *Elmis sp.* is a bioindicator of clean to mildly polluted waters.

Not all of the induced rules agree with existing expert knowledge. As an example, let us consider the rules that predict the presence of the taxon *Plecoptera leuctra sp. (53000)*, which is used as an indicator of clean waters. The induced rules do confirm that it is found mainly in clean waters. However, they also state that *Plecoptera leuctra sp.* can be found in quite polluted water, provided there is enough oxygen. Thus, they enhance current knowledge on the bioindicator role of this taxon.

Another example are the rules that predict the presence of the taxon *Cymbella sp. (13200)*. The rules point out that the taxon can be found in moderately to critically polluted waters (as indicated by the tolerance of large quantities of biodegradable matter, i.e., large values of *BOD*). In water monitoring practice, however, *Cymbella sp.* is used as an indicator of clean to mildly polluted waters.

Biological classification of river water quality

This section describes the experiments in predicting the biological water quality class, as determined by the saprobic index, from two different sets of attributes. The first set consists of all the physical and chemical parameters mentioned in the previous section, altogether 13 parameters. The second consists of the 17 taxa from the previous section and 10 additional taxa, altogether 27 taxa. The 13 parameters give rise to real valued attributes, while the 27 taxa give rise to discrete valued attributes with four (linearly ordered) values: 0, 1, 3, and 5. As mentioned in the introduction, there are seven water quality classes. The majority class 2. comprises 339 of the 698 examples, thus the default accuracy is 48.6%.

```
IF  Temperature < 12.55
AND PH < 8.45
AND NO2 < 0.235
AND 1.75 < NO3 < 7.15
AND Detergents < 0.025
AND COD < 4.25
AND BOD < 2.35
THEN QualityClass = 1.-2.
     [9 80 2 0 0 0 0]
```

```
IF Temperature > 12.65
AND PH < 8.65
AND Saturation > 57.3
AND NO2 < 0.375
AND NH4 > 0.065
AND PO4 < 0.39
AND SiO2 < 10.75
AND COD > 2.65
AND 1.25 < BOD < 4.75
THEN QualityClass = 2.
     [0 8 152 5 0 0 0]
```

For illustration, let us take a look at two rules that predict the water qualy class from physical and chemical parameters: these are the rules with greatest coverage derived from the whole dataset. While we would need expertise in both the chemistry and biology of water quality to thoroughly evaluate these rules, they are intuitive and understandable. The class 1.-2. requires relatively cold water and very small quantities of pollutants (NO_2, NO_3, detergents, *COD*, *BOD*). Class 2. waters are usually warmer and somewhat larger quantities of pollutants are allowed, provided there is enough oxygen (`Saturation > 57.3`).

The rules induced on the entire dataset reach 81.5% classification accuracy when using physical and chemical parameters and 71.1% when using bioindicators, the information scores being 62% and 44%. When learning on 70% of the dataset, the corresponding accuracies on the testing set are 60% and 58%, the information scores being 32% and 28%. It is interesting to note that better performance is achieved when predicting from physical and chemical parameters, despite the fact that biological quality is predicted. However, to determine the quality class a much larger set of bioindicators is used than the one used in our experiments.

```
IF  BACILLARIOPHYTA_Navicula_cryptocephala = 0
AND CHLOROPHYTA_Scenedesmus_obliquus = 0
AND DIPTERA_Chironomidae_green = 3
AND COLEOPTERA_Elmis_sp. = 3
THEN QualityClass = 1.-2. [1 16 1 0 0 0 0]
```

```
IF  BACILLARIOPHYTA_Navicula_cryptocephala = 1
AND BACILLARIOPHYTA_Nitzschia_palea = 1
THEN QualityClass = 2. [0 3 32 9 2 0 1]
```

Let us finally take a look at the two rules above that predict the water quality class from the 27 bioindicators. The first predicts class 1.-2. when *Elmis sp. (59500)* and *Chironomidae green (58900)* occur frequently (3) and the species *Scenedesmus obliquus (31900)* and *Navicula cryptocephala (17700)* are absent (0). It is in agreement with existing expert knowledge: *Elmis sp.* and *Chironomidae green* are indicators of clean waters, while *Navicula cryptocephala* is indicative of polluted waters (class 3.). The second rule predicts class 2. when *Navicula cryptocephala* and *Nitzschia palea* occur incidentally (1). Both species are indicative of heavily polluted waters if they occur in larger quantities: as they only occur incidentally, the rule is in agreement with expert knowledge.

Discussion

The experiments we have performed indicate that rule induction can be used to analyze water quality data and discover different kinds of knowledge. We induced rules that describe the influence of physical and chemical properties of the water in Slovenian rivers on the presence of selected living organisms that are currently used as bioindicators of river water quality. Expert evaluation of these rules showed that they do indeed capture useful knowledge, as indicated by their positive information scores. In some cases, the rules just confirmed the expert knowledge of the biology of the bioindicator taxon concerned. In others, it revealed new aspects of the biology of the studied taxon, which extend existing knowledge without conflicting it. There were even cases when the rules indicated that the given taxon is used as an indicator for a wrong class of biological water quality.

While the above analysis concerned 17 taxa with relatively well known biology that are routinely used as bioindicators of water quality, we have still been able to find some new knowledge on the biology of the taxa studied and their bioindicator roles. A promising direction for further work is to extend the analysis to taxa about which relatively little is known and which are currently not used as bioindicators. This could contribute new knowledge both to biology and to the practice of water quality monitoring, as some of the taxa analyzed may turn out to be very good bioindicators. Additional methods of data analysis from the areas of machine learning and statistics may be used in the process.

We also induced rules that predict the river water quality class (as provided through the saprobic index). The rules that use bioindicator data to this end are mainly consistent with existing expert knowledge: this is understandable, as bioindicator data (albeit on a much larger set of indicators) is used to derive the saprobic index. The rules that predict the biological quality class from the physical and chemical water properties are surprisingly accurate and informative and deserve a more detailed further analysis by experts fluent in both biological and chemical aspects of water quality. It would be reasonable to induce classification rules that use both bioindicator and chemical/physical data, as the two are complementary to a certain degree.

A serious problem is the use of the saprobic index for classification. The saprobic index and similar methods are rather ad hoc and leave much to be desired (Walley, Boyd, & Hawkes 1992). An alternative approach that has been tried for British rivers is to have an expert classify the samples and then induce classification rules (Walley, Boyd, & Hawkes 1992; Džeroski et al. 1994). Another possibility is to use clustering methods (with a heavy expert interaction) to identify the classes and then induce rules that describe these classes. We hope to investigate the success of these approaches for Slovenian rivers and propose a more objective and reliable methodology for water quality assessment.

Acknowledgements

We would like to thank Doris Ličan-Milošević, who performed the experiments described here as a part of her BSc Thesis at the Faculty of electrical engineering and computer science, University of Ljubljana, Slovenia. Her thesis was supervised by Professor Ivan Bratko and the authors. We acknowledge the support of the Hydrometeorological Institute of Slovenia that provided the water quality data used in the experiments, as well as the support of the Slovenian Ministry of Science and Technology.

References

Cestnik, B. 1990. Estimating probabilities: A crucial task in machine learning. In Proceedings of the Ninth European Conference on Artificial Intelligence, 147–149. London: Pitman.

Clark, P. and Boswell, R. 1991. Rule induction with CN2: Some recent improvements. In Proceedings of the Fifth European Working Session on Learning, 151–163. Berlin: Springer.

De Pauw, N. and Hawkes, H. 1993. Biological monitoring of river water quality. In Proceedings of the Freshwater Europe Symposium on River Water Quality Monitoring and Control, 87–111. Birmingham: Aston University.

Džeroski, S., Cestnik, B., and Petrovski, I. 1993. Using the *m*-estimate in rule induction. *Journal of Computing and Information Technology* 1:37–46.

Džeroski, S., Dehaspe, L., Ruck, B., and Walley, W. 1994. Classification of river water quality using machine learning. In Proceedings of the Fifth International Conference on the Development and Application of Computer Techniques to Environmental Studies, volume I, 129–137. Southampton: Computational Mechanics Publications.

Grbović, J. 1994. Applicability of Various Procedures for the Assessment of Quality of Torrential Streams. PhD Thesis, Biotechnical Faculty, University of Ljubljana, Slovenia. In Slovenian.

Kolkwitz, R. and Marsson, M. 1902. Grundzatze für die biologische Beurteilung des Wassers nach seiner Flora und Fauna. *Mitt. Prüfungsanst. Wasserversorg. Äbwasserein* 1:33–72.

Kononenko, I. and Bratko, I. 1991. Information-based evaluation criterion for classifier's performance. *Machine Learning* 6:67–80.

Ličan-Milošević, D. 1994. Analysis of water quality data by rule induction. BSc Thesis, Faculty of electrical engineering and computer science, University of Ljubljana, Slovenia. In Slovenian.

Official, J. 1978. Regulations on water classification for interstate streams, international waters and the coastal waters of Yugoslavia. *Official Journal of the SFRY*, Number 6.

Pantle, R. and Buck, H. 1955. Die biologische Überwachung der Gewas und die Darstellung der Ergebnisse. *Gas und Wasserfach* 96:603.

Walley, W., Boyd, M., and Hawkes, H. (1992). An expert system for the biological monitoring of river pollution. In Proceedings of the Fourth International Conference on the Development and Application of Computer Techniques to Environmental Studies, 1030–1047. Amsterdam: Elsevier.

Zelinka, M. and Marvan, P. 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit fliessender Gewässer. *Arch. Hydrobiol.* 57:389–407.