Aquat Ecol (2007) 41:399–411
DOI 10.1007/s10452-006-9058-y

ORIGINAL PAPER

# Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece)

**Eleni Dakou · Tom D'heygere · Andy P. Dedecker · Peter L. M. Goethals · Maria Lazaridou-Dimitriadou · Niels De Pauw**

Received: 30 May 2006 / Accepted: 2 June 2006 / Published online: 7 July 2006


**Abstract** In this study, decision tree models were induced to predict the habitat suitability of six macroinvertebrate taxa: Asellidae, Baetidae, Caenidae, Gammaridae, Gomphidae and Heptageniidae. The modelling techniques were applied on a dataset of 102 samples collected in 31 sites along the river Axios in Northern Greece. The database consisted of eight physical-chemical and seven structural variables, as well as the abundances of 90 macroinvertebrate taxa. A seasonal variable was included allowing the description of potential temporal changes in the macroinvertebrate taxa. Rules relating the presence/absence of six benthic macroinvertebrate taxa with the 15 physical-chemical and structural river characteristics and the seasonal variable were induced using the J48 algorithm. In order to improve the performance and the interpretability of the induced models, three optimisation techniques were applied: tree-pruning, bagging and boosting. The predictive performance of the decision tree models was assessed on the basis of the percentage of Correctly Classified Instances (CCI) and the Cohen's kappa statistic. The results of the present study demonstrated that although the models had a relatively high predictive performance, noise in the dataset and inappropriate input variables prevented to some extent, the models from making reliable predictions. Although tree-pruning did not improve significantly the reliability of the induced models, it reduced considerably the tree complexity and in this way increased the transparency of the trees. Consequently, the induced models allowed for a correct ecological interpretation. The effect of bagging and boosting on the other hand varied considerably between the different models, as well as within different repetitions of 10-fold cross-validation in an individual model. In some cases the predictive performance was improved, in others stable or even worsened. The effect of bagging and boosting seemed to be strongly dependent on the dataset on which the two techniques were applied. Tree-pruning thus proved to have a high potential when applied in models used for decision-making of river restoration and conservation management.

**Keywords** Bagging · Biomonitoring · Boosting · Decision tree · Ecological modelling · Habitat suitability · River assessment · Tree-pruning

E. Dakou · M. Lazaridou-Dimitriadou
Laboratory of Zoology, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece

T. D'heygere · A. P. Dedecker · P. L. M. Goethals (✉) · N. De Pauw · E. Dakou
Department of Applied Ecology and Environmental Biology, Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, J. Plateaustraat 22, B-9000 Ghent, Belgium
e-mail: peter.goethals@UGent.be

## Introduction

Due to point source pollution of untreated urban and industrial wastewater and diffuse pollution originating from agricultural activities, the ecological quality of water bodies in Greece has gradually been decreasing during the last decades. Measures to halt this degradation and restore the waters thus become more and more a necessity.

Knowledge about the relationship between the environmental factors and the occurrence of freshwater organisms is a key issue in conservation management and river restoration. Assessment of sites that could support important taxa, or prediction of the responses of target species on changes of land use or river structure are a few examples for which the insight in the species–environment relationship is needed. In this context, modelling is becoming an essential tool to support decision-making in water management. Modelling of river ecosystems for instance has made substantial progress during recent years. Nevertheless, the non-linear and complex nature of ecosystems makes this understanding still difficult and only a gradual progress in adequate ecosystem modelling and computation has been obtained (Recknagel 2002). The availability of proper datasets and modelling techniques, however, now seems to allow for the development of ecosystem models with a high reliability. Recently, new concepts are being more commonly used to analyse ecosystem databases and to make predictions for river management purposes. Artificial neural networks (Lek and Guegan 1999), fuzzy logic (Barros et al. 2000), decision trees (Quinlan 1986) and Bayesian belief networks (Adriaenssens et al. 2004) for instance have proven to have a high potential in habitat suitability modelling, as they combine reliable predictions with a convenient interpretation of the predictive results (Goethals and De Pauw 2001; Goethals 2005).

One well-studied data soft-computing method, induction of classification and regression trees (often referred to as decision trees when discussing both methods) has been shown to be useful in modelling complex datasets (Breiman et al. 1984). In contrast to ANNs, the application of classification and regression trees in ecological modelling, in particular related to macroinvertebrates, is rather limited and hardly described in literature (Goethals 2005). In the following paragraph, an overview of the major examples is presented.

Kompare et al. (1994) described some general possibilities of machine learning in the field of ecology. Dzeroski et al. (1997) were among the first to describe applications of classification trees in river community analysis. These include the biological classification of British rivers based on bioindicator data, the analysis of the influence of physical and chemical parameters on selected bioindicator organisms in Slovenian rivers and the biological classification of Slovenian rivers based on physical and chemical parameters as well as bioindicator data. In all three cases, valuable models (knowledge) in the form of rules were extracted from data acquired through environmental monitoring and/or expert interpretation of the acquired samples. The applied algorithm was CN2 (Clark and Niblett 1989). H. Blockeel et al. (1999a unpublished) applied TILDE to predict an ecological index (Saprobic Index) for Slovenian rivers. The input variables were biological data, physical-chemical characteristics (actual and time-series) as well as combinations. Additionally, also macroinvertebrate communities were successfully predicted on the basis of physical-chemical variables. In H. Blockeel et al. (1999b unpublished), physical-chemical variables were predicted on the basis of biological communities. Innovative in this article is the use of a single tree to predict all these variables at once, what eases the use of this relative simple information in river management. In Dzeroski et al. (2000), the prediction of physical-chemical variables was established on the basis of biological data. The research revealed that certain taxa occurred in many trees, what makes them useful to be selected as indicator taxa. The research proved as well that when compared to linear regression, the model seemed to give the same performance. Dzeroski and Drumm (2003) applied regression trees (M5′) programme (a Java implementation of the M5 algorithm in WEKA (Witten and Frank 2000)), to predict sea cucumbers (*Holothuria leucospilota*) in lagoons around the Cook Islands. Based on these trees they were able to retrieve

the preferred habitat of this species and found out that the dominant variables are rubble and sand.

The aim of this paper is to demonstrate the potential and limitations of decision trees as a habitat suitability model for macroinvertebrates in the river Axios in Northern Greece. Three optimisation techniques for improving the accuracy and transparency of the models, namely tree-pruning, bagging and boosting, have been examined for their potential use in river management.

## Materials and methods

### Study area

The river Axios originates in the Sar Mountains of the Former Yugoslavian Republic of Macedonia (FYROM) (Fig. 1). It discharges into the Thermaikos Gulf in northern Greece. Only the last 80 km of the 320-km long river are within Greek territory. At 49 km from the border with FYROM an irrigation dam (Fragma Ellis) has been constructed, which remains closed from May to September. Due to this fact, discharge falls to 1 $m^3/s$ during the dry season (Argiropoulos 1991). Agriculture is the dominant land use activity within the watershed and is, as a consequence, the main source of pollution. Next to this diffuse pollution, urban and industrial wastes are being discharged in several places along the river (Fig. 1).

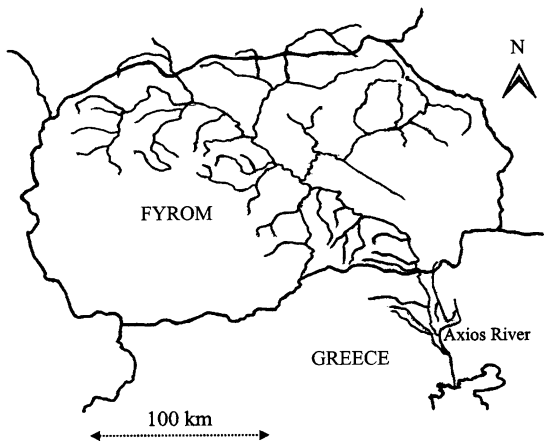


**Fig. 1** The location of the Axios river in FYROM and Greece

### Measurements and database set-up

The database consisted of measurements of 102 samples originating from the Axios River, collected in 31 sites located in the Greek part between 1997 and 2001. In some sites, the measurements were conducted several times a year (between 2 and 12 measurements) to be able to detect seasonal changes in the macroinvertebrate communities, while other sites were sampled only once between 1997 and 2001. Nineteen environmental variables were measured during the monitoring campaigns. Further information on the selection of the sampling sites and the sampling methodologies is given by Langrick et al. (1998), Kampa et al. (2000) and Chatzinikolaou (2001, 2002). Of those 19 variables, 15 were used as inputs to predict the habitat suitability of the macroinvertebrate taxa (Table 1). The applied input variables included information on the physical-chemical as well as the structural characteristics of the river. Also a variable was introduced to account for the seasonal variation in the dataset.

The samples of benthic macroinvertebrates were taken by means of the 3-min kick-sweep method (Armitage et al. 1983), using a standard pond net (surface 575 $cm^2$, mesh size 900 μm, depth 27.5 cm). Identification of the macroinvertebrates was carried out to family level, in accordance with the Greek Biotic Index (V. Artemiadou and M. Lazaridou-Dimitriadou submitted). In total, 90 taxa were found of which 6 were selected for modelling their habitat suitability. The selection of the macroinvertebrate taxa was based on their frequency of occurrence, as well as on their sensitivity to pollution. The sensitivity to pollution according to the Greek Biotic Index (100 = high sensitivity, 10 = high tolerance to pollution) (V. Artemiadou and M. Lazaridou-Dimitriadou submitted) of the 6 selected taxa is presented in Table 2. In the models, the absence or presence of macroinvertebrate taxa was represented respectively by 0 or 1.

### Rule induction with the J48 algorithm

The common way to induce rules in the form of decision trees is the so-called ‘‘Top-Down

**Table 1** Input variables used for the prediction of the habitat suitability of six macroinvertebrate taxa in the Axios River in Northern Greece and their respective minimum and maximum values

| Variables | Minimum value | Maximum value | Units |
|---|---|---|---|
| pH | 7.0 | 9.7 | – |
| Dissolved Oxygen Saturation (%DO) | 5.0 | 174.8 | % sat |
| Dissolved Oxygen (DO) | 0.1 | 17.4 | mg/l |
| Biological Oxygen Demand ($BOD_5$) | – 1.1 | 15.4 | mg/l |
| Temperature | 4.0 | 30.2 | °C |
| Conductivity | 178 | 4,860 | $\mu$S/cm |
| Total Dissolved Solids (TDS) | 80 | 2,430 | mg/l |
| Total Suspended Solids (TSS) | 0 | 194 | mg/l |
| Flow velocity | 0.0 | 5.9 | m/s |
| Granulometric classification of substrate | | | |
| Boulders | 0 | 85 | |
| Cobbles | 0 | 90 | |
| Pebbles | 0 | 50 | |
| Gravel | 0 | 65 | |
| Sand | 0 | 95 | |
| Silt | 0 | 100 | % |
| Season | | | 4 classes (Winter, Spring, Summer, Autumn) |

Induction of Decision Trees'' (Quinlan 1986). Tree construction proceeds recursively, starting with the entire set of training examples. For each step, the most informative input variable is selected as the root of the sub-tree and the current training set is split into subsets according to the values of the selected input variable. In this manner, rules are generated that relate the values of input variables with the presence/absence of macroinvertebrate taxa. For discrete input variables, a branch of a tree is typically created for each possible value of that particular variable. For continuous input variables, a threshold is selected and two branches are created based on that threshold. Tree construction stops when all examples in a node are of the same class (or if some other stopping criterion is satisfied). Such nodes are called leaves and are labelled with the corresponding values of the class.

The C4.5 algorithm (Quinlan 1993) is one of the most well-known and widely used decision tree induction method. The J48 algorithm is a Java re-implementation of C4.5 and is a part of the machine-learning package WEKA (Witten and Frank 2000). In the following experiments, the J48 algorithm with binary splits was used for the induction of classification trees. Binary split is a parameter of the J48 algorithm that decides whether a node can only split into two branches or more.

The input variables of the models consisted of physical-chemical and structural measurements, some of which were continuous and others discrete, while the output variables were macroinvertebrate taxa, which were discrete (presence or absence).

**Table 2** Selected macroinvertebrate families used for the modelling approach and their respective sensitivity to pollution according to the Greek Biotic Index (V. Artimiadou and M. Lazaridou-Dimitriadou submitted)

| Taxon | Frequency of occurrence (%) | Sensitivity to pollution |
|---|---|---|
| Asellidae (Crustacea, Isopoda) | 34 | 30 |
| Baetidae (Insecta, Ephemeroptera) | 69 | 40 |
| Caenidae (Insecta, Ephemeroptera) | 61 | 50 |
| Gammaridae (Crustacea, Amphypoda) | 41 | 50 |
| Gomphidae (Insecta, Odonata) | 54 | 60 |
| Heptageniidae (Insecta, Ephemeroptera) | 9 | 80 |

## Optimisation techniques

In order to reduce the noise in the data and to improve the predictive results with regard to complexity and accuracy of the predictions, three optimisation methods were applied: tree-pruning, bagging and boosting. A common way to cope with tree complexity is tree-pruning. Optimal tree-pruning is an important mechanism as it improves the transparency of the induced trees by reducing their size, as well as enhances their classification accuracy by eliminating errors that are present due to noise in the data (Bratko 1989). There are two types of tree-pruning: forward pruning and post-pruning. When forward pruning is applied, the expansion of the tree is stopped when a certain criterion is met. For example, every leave should contain a minimum number of instances or no branching is allowed. Post-pruning on the other hand, means that first a highly branched tree is constructed. Afterwards, some of the ending subtrees are replaced by leaves based on their reliability. The reliability of the subtrees is evaluated by comparing the classification error estimates before and after replacing a subtree by a leave. In the following experiments, post-pruning was used. By changing the confidence factor ($c$), the intensity of pruning was controlled. The confidence factor is a parameter that has an effect on the error rate estimate in each node. When the confidence factor is increased, the difference between the error estimate of a parent node and its splits decreases. In this way, it is less likely that the split will be pruned. The smaller the value of the confidence factor is, the larger is the difference between the error rate estimates of a parent node and its potential splits. Thus, the chance that splits will be replaced by leaves is higher.

Bagging (Bootstrap aggregation) and AdaBoost (Adaptive Boosting) (Witten and Frank 2000) are voting classification algorithms. Bagging and boosting are used in combination with the base classifier that creates 'child' datasets from a single 'parent' dataset that is originally used for training. This allows for taking advantage of the inherent instability of the base classifier. The instability of a classifier is defined as the tendency to find large changes in the predicted values caused by minor changes in the dataset (Breiman 1996). In bagging, the 'child' datasets are created by duplicating some of the instances of the 'parent' dataset randomly and deleting others. From each 'child' dataset, a different tree is constructed that leads to a different prediction. The different predictions of the 'child' datasets are combined by a majority vote to give the final prediction. Boosting also creates 'child' datasets from a single 'parent' dataset, but the difference is that each new 'child' dataset is influenced by the previous one, as the instances that are duplicated are not randomly selected. The instances that are incorrectly predicted in a dataset are included in the next dataset as duplicated ones, so that the chance of a correct prediction of these previously misclassified instances improves. These duplicated instances will affect the training of the model and therefore also the resulting classification tree. This procedure continues until a pre-defined number of iterations is reached, but it stops earlier in case the error estimate is lower than 0.05. In the following experiments bagging and boosting were applied on the J48 algorithm, which included binary splits and optimal pruning. The two techniques are included in the machine-learning package WEKA. Both techniques included ten iterations, while the size of the training datasets created by the two algorithms was the same as the original training dataset.

## Model training and validation

The predictive models were evaluated on the basis of two performance measures. This required the derivation of matrices of confusion from the modelling results that identified true positive (TP), false positive (FP), false negative (FN) and true negative (TN) cases predicted by each model (Fielding and Bell 1997) (Table 3). In this way, the presence/absence patterns were tabulated against those predicted.

The first performance measure that was calculated was the percentage of Correctly Classified Instances (CCI):

$$\mathrm{CCI} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})} \times 100$$

**Table 3** The derivation of the confusion matrix used as a basis of performance measures in presence/absence models with true positive (TP), false positive (FP), false negative (FN) and true negative values (TN)

| | Actual | |
|---|---|---|
| | Present | Absent |
| Predicted | | |
| Present | TP | FP |
| Absent | FN | TN |

Another performance measure that was calculated was the Cohen's kappa statistic (Cohen 1960). It is a derived statistic that measures the proportion of all possible cases of presence or absence that are predicted correctly by a model after accounting for chance predictions. It is calculated as:

$$\text{Kappa} = \frac{[(\text{TP}+\text{TN}) - (((\text{TP}+\text{FN})(\text{TP}+\text{FP}) + (\text{FP}+\text{TN})(\text{FN}+\text{TN}))/n)]}{[n - (((\text{TP}+\text{FN})(\text{TP}+\text{FP}) + (\text{FP}+\text{TN})(\text{FN}+\text{TN}))/n]}$$

Landis and Koch (1977) attempted to indicate the degree of agreement that exists when the Cohen's kappa is found to be in various ranges: ≤ 0 (poor); 0–0.2 (slight); 0.2–0.4 (fair); 0.4–0.6 (moderate); 0.6–0.8 (substantial); 0.8–1 (almost perfect).

Model training and validation was based on a stratified 10-fold cross-validation (Kohavi 1995). To allow a reliable error estimate of the models, ten stratified 10-fold cross-validation experiments were carried out, from which the average predictive performance was calculated. These ten-time repeated 10-fold cross-validation experiments allowed for determining the 95% confidence limit of the average predictive performance, considering that the data were normally distributed. This confidence limit is represented in the tables and figures in the results.

For the comparison of the performance of the different applied techniques a paired Student's *t*-test was performed (Witten and Frank 2000). The results of the same partitions of the repeated 10-fold cross-validations were compared. For a two-tailed test, a significance level of 5% was used.

## Results

### Model development and validation

In this study, models for the prediction of the habitat suitability of six macroinvertebrate taxa were induced by using the J48 algorithm with binary splits. The predictive results of the constructed unpruned trees are presented in Table 4. The average CCI (%) and Cohen's kappa statistic of the repeated ten 10-fold cross-validations with a 95% confidence interval of the average are presented, as well as the number of leaves of each tree. The percentage of CCI was in all cases relatively high. Cohen's kappa statistic was high in the models for the prediction of Gomphidae, and relatively high in the case of Caenidae, while it had low values in all the other cases. A value of Cohen's kappa statistic above 0.4 is considered to indicate a reliable model, while lower values indicate poor model performance. The low values of Cohen's kappa statistic in the induced models revealed that most of the predictions were based on chance and especially the predictive model of Heptageniidae, for which a kappa value of 0.108 was found. Additionally, the constructed trees in

**Table 4** Predictive results of models based on the J48 algorithm without pruning optimisation

| Taxon | CCI (%) | Cohen's kappa | Number of leaves |
|---|---|---|---|
| Asellidae | 66.3 ± 2.0 | 0.247 ± 0.04 | 18 |
| Baetidae | 68.1 ± 1.7 | 0.278 ± 0.04 | 17 |
| Caenidae | 66.8 ± 2.1 | 0.306 ± 0.04 | 14 |
| Gammaridae | 64.3 ± 2.2 | 0.260 ± 0.04 | 16 |
| Gomphidae | 79.3 ± 1.8 | 0.583 ± 0.04 | 14 |
| Heptageniidae | 85.6 ± 2.0 | 0.108 ± 0.06 | 6 |

**Table 5** Predictive results of models induced on the basis of the J48 algorithm by using pruning optimisation

| Taxon | CCI (%) | Cohen's kappa | Number of leaves | Confidence level |
|---|---|---|---|---|
| Asellidae | 68.3 ± 1.5 | 0.278 ± 0.03 | 14 | 0.25 |
| Baetidae | 72.5 ± 1.9 | 0.320 ± 0.04 | 12 | 0.15 |
| Caenidae | 70.0 ± 1.7 | 0.351 ± 0.04 | 3 | 0.15 |
| Gammaridae | 62.2 ± 1.3 | 0.218 ± 0.03 | 10 | 0.15 |
| Gomphidae | 78.8 ± 2.0 | 0.573 ± 0.04 | 11 | 0.25 |
| Heptageniidae | 90.2 ± 0.50 | − 0.017 ± 0.01 | 1 | 0.15 |

most cases included a large amount of leaves, which increased their complexity and prevented an ecological interpretation.

## Pruning optimisation

In order to reduce the complexity of the constructed trees and improve the predictive performance of the models, tree-pruning was performed. Models with different intensity of pruning were induced by varying the confidence factor between 0.15 and 0.25. The optimal confidence level of pruning was different for models predicting the habitat suitability of different taxa. Paired Student's *t*-tests were conducted for all the induced models in order to compare the average CCI (%) and Cohen's kappa statistic over ten 10-fold cross-validations before and after pruning. In Table 5 the predictive results of the pruned trees are presented. A significant increase in the predictive performance of pruned trees based on the CCI (%) was detected for all the taxa except for Gammaridae and Gomphidae. For both taxa, the predictive performance decreased. However, the decrease was not significant according to the results of the Student's *t*-test. Cohen's kappa statistic increased significantly for Asellidae and Baetidae, while it decreased significantly for Heptageniidae. Although the CCI (%) of Heptageniidae increased, Cohen's kappa was reduced to a value that revealed a complete unreliability of the model, while at the same time a failure of the model to construct a tree was observed. This is indicated by the fact that the tree consisted of only one leaf. Although kappa statistic improved for most of the species, still only one reliable model was constructed. Also the complexity of the constructed trees was reduced considerably. The highest decrease in the number of leaves was found for Caenidae.

## Bagging and boosting optimisation

In an attempt to further optimise the predictive models, bagging and boosting were applied on the J48 algorithm that already included binary splits and pruning. Paired Student's *t*-tests were conducted for the comparison of the predictive performance of models based on the J48 algorithm with and without bagging and boosting. The effect of bagging and boosting on the predictive performance of the models are presented in Figs. 2 and 3. In Fig. 2, the predictive performance is estimated by the average CCI (%) of the ten repeated 10-fold cross-validations with a confidence level of 95%. On the other hand, in Fig. 3 the predictive performance is estimated by the average of the Cohen's kappa statistic of the ten repeated 10-fold cross-validations with a confidence level of 95%. When bagging was applied, a statistically significant increase in the predictive performance was detected for Asellidae and Baetidae, while the performance of the other models either increased or decreased but never significantly. Boosting had a significant effect on the predictive performance of Asellidae, Caenide, Gammaridae and Heptageniidae. In the case of Heptageniidae, the effect of boosting referred only to Cohen's kappa statistic, while it did not significantly affect the CCI (%). It could be observed from Fig. 2 that the effect of the two techniques varied in relation to the organism predicted and did not follow a general trend. Thus, in the cases of Asellidae and Gammaridae, the two techniques resulted in improved predictions, while in the case of Caenidae and Gomhidae, a reduced predictive performance was found. In the case of Heptageniidae, bagging improved the predictive results, while boosting reduced the overall performance. The effect of bagging and boosting on the model performance based on

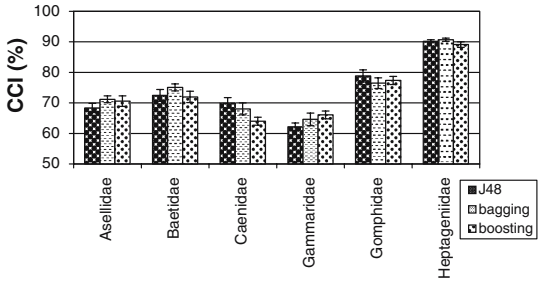


**Fig. 2** Assessment of bagging and boosting based on the percentage of Correctly Classified Instances (CCI %) for the six selected macroinvertebrate taxa in the Axios river (Northern Greece)

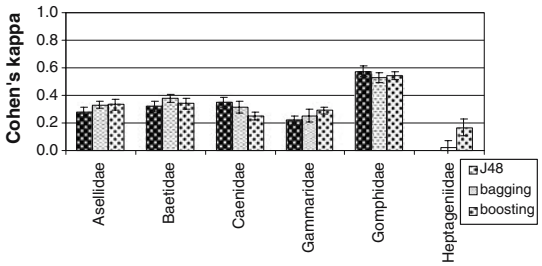


**Fig. 3** Assessment of bagging and boosting based on Cohen's kappa statistic for the six selected macroinvertebrate taxa in the Axios river (Northern Greece)

Cohen's kappa statistic was similar to the one on the performance based on the CCI (%). However, there was an exception. It was observed that boosting led to a significant improvement of Cohen's kappa statistic, although it did not improve the CCI (%) in the predictive model for Heptageniidae. Nevertheless, the model could be considered as irrelevant because the Cohen's kappa was still lower than 0.4.

In order to obtain a better insight in the way the two techniques function, the effect of bagging and boosting on the percentage of correctly predicted presence and absence cases in each model was examined. Paired Student's *t*-tests were conducted for the comparison of the percentage of correctly predicted presence and absence cases, as well as for the overall CCI (%) with and without the application of bagging and boosting. It was observed that both bagging and boosting lead to a significant increase of the percentage of correctly predicted absence cases. A small decrease of the percentage of correctly predicted presence cases was also observed. Bagging did not seem to significantly alter the overall CCI (%), while boosting increased it significantly. In Fig. 4, the effect of bagging and boosting on the predictions of presence cases, absence cases, together with the overall predictive capacity for Gammaridae is presented. It was obvious that although the two techniques did not have a considerable effect on the overall percentage of CCI, there was a significant increase of the percentage of correct predictions as absent and a decrease of the percentage of correct predictions as present. This phenomenon became more pronounced when bagging was applied.

In most cases in which bagging and boosting was applied, an increase in the variability of the predictive results between the different repetitions of 10-fold cross-validation was observed, as well as an increase on the 95% confidence intervals of the average performances. Thus, the effect of the two techniques on the different partitions of the dataset that were used for the repetition of the stratified 10-fold cross-validation, could be observed. In Fig. 5, the effect of bagging and boosting on two stratified 10-fold cross-validations based on different partitioning of the same dataset is presented. The model predicts the habitat suitability for Gammaridae. In Fig. 5a, it is observed that bagging and boosting caused an increase on the correct predictions made as present, and a decrease on the correct predictions made as absent, while the overall CCI (%) was relatively stable. In Fig. 5b, bagging and boosting had exactly the opposite effect, an increase of the correct predictions made as absent, and a decrease of the correct predictions made as present. The overall CCI (%) also remained stable. The

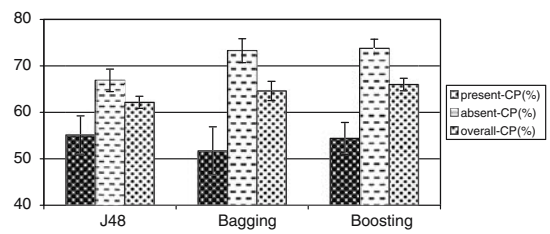


**Fig. 4** The effect of bagging and boosting on the percentage of Correctly Classified Instances of presence cases, absence cases and on the overall CCI (%) for the prediction of Gammaridae in the Axios river (Northern Greece) (CP = correctly predicted)

contradictive predictive results in these two cases revealed that although there was a general trend of the effect of bagging and boosting on the predictive results, which is indicated by the average performance of ten repeated 10-fold cross-validations, the application of the two techniques could lead to predictive results that vary not only in relation to the predicted organism, but also within an individual model. A strong dependency between the training dataset and the effect of bagging and boosting was detected, as different partitioning of the same dataset could lead to opposite results.

In Fig. 6, the effect of bagging and boosting on the predictions of presence cases, absence cases and on the overall predictive capacity for Heptageniidae is presented after ten 10-fold cross-validations. It is observed that although the two techniques did not have a considerable effect on the overall percentage of CCI, there was a significant increase of the percentage of correct predictions as present that reaches 17.8% and a decrease of the percentage of correct predictions as absent. This phenomenon became more obvious after the application of boosting. However, the variation of the predictions of the presence cases within the ten repetitions of stratified 10-fold cross-validation, was very high and seemed to increase as the percentage of correct predictions was increasing. The frequency of occurrence of Heptageniidae in the dataset used was very low and an examination of the confusion matrices of the induced models for the prediction of this taxon revealed that in most cases the models failed to make any correct prediction of the presence cases. When the J48 algorithm was applied without a combination of bagging or boosting, the algorithm was not able to make even one correct prediction of the presence cases. When bagging was applied, one of the nine presence cases was predicted correctly. When boosting was applied, maximum two presence cases were correctly predicted. Thus, the 17.8% increase of the percentage of correctly predicted presence cases equalled with only two cases correctly predicted. This last statement underlines that even when bagging and boosting seem to perform well at first sight, a thorough examination of their effects is needed before applying these techniques for management purposes.

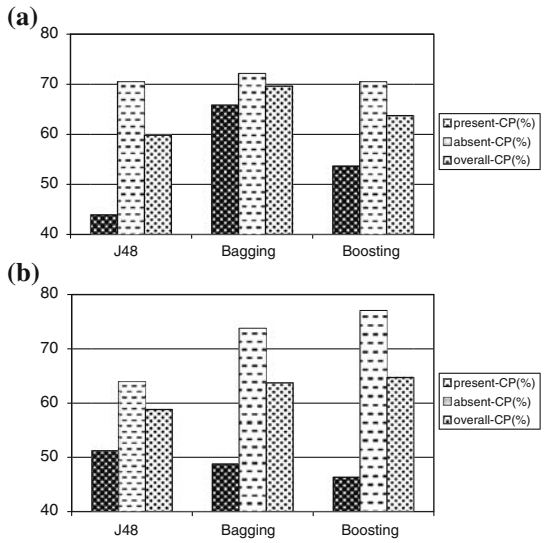


**Fig. 5** The effect of bagging and boosting on the percentage of Correctly Classified Instances as present or absent and on the overall CCI (%) for the prediction of Gammaridae in the river Axios (Northern Greece). Two repetitions of 10-fold cross-validation based on two different partitions of the dataset are presented in (**a**) and (**b**) (CP = correctly predicted)

## Discussion

### Model development and validation

When using the CCI (%) as an evaluation measure, decision trees generally performed well to predict the habitat suitability of the six selected macroinvertebrate taxa in the Axios river. However, the

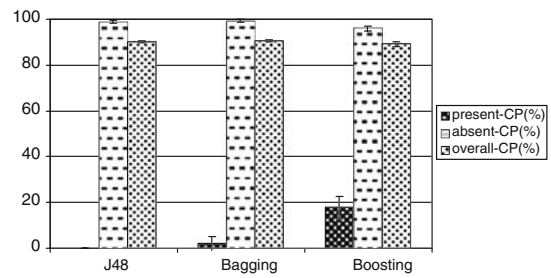


**Fig. 6** The effect of bagging and boosting on the percentage of Correctly Classified Instances of presence cases, absence cases and on the overall CCI (%) for the prediction of Heptageniidae in the Axios river (Northern Greece) (CP = correctly predicted)

low value of the Cohen's kappa statistic revealed that most of the models did not yield reliable predictions, as they were mainly based on chance. Noise in the data, missing values in the dataset, as well as the fact that unmeasured input variables could explain the habitat suitability of the predicted taxa, are probably the main reasons for the incapability of some models to make reliable predictions. In a study on similar models in Flanders, D'heygere et al. (2003) demonstrated that some of the most important input variables for the prediction of the presence or absence of macroinvertebrates were depth, Kjeldahl nitrogen and ecotoxicological variables. The introduction of some structural, physical-chemical and even ecotoxicological measurements in the dataset, could therefore improve the predictive capacity of the induced models. Furthermore, a selection of the most appropriate input variables could be useful for an improvement of the predictive performance of the models (D'heygere et al. 2003). A major disadvantage of the induced models was the complexity of the constructed trees. In most cases, the constructed trees consisted of a large amount of leaves. In this way, the detection of general trends in the data was very difficult. Thus, the ecological interpretation of the results was not possible, as the trees were unable to offer information about the habitat suitability for the macroinvertebrate taxa. The fact that the minimum number of instances in a leaf was two, in addition to the criterion used for stopping tree expansion, lead to the construction of very detailed trees, in which even individual cases were presented. As a consequence, the ability to generalise in the models is being reduced.

## Pruning optimisation

The application of tree-pruning resulted in relatively simple and more understandable trees that could be ecologically interpreted. The confidence factor that produced optimal results with regard to the CCI (%) and Cohen's kappa, varied in relation to the predicted taxon. Thus, it seems that in each model the intensity depends on the data itself. Tree-pruning resulted in a significant increase of the CCI (%). This increase could be an indication of the improvement of the accuracy of the models for some taxa. As the noise of the data has a reduced effect on the models after pruning, such an improvement was expected. However, the effect of pruning on the Cohen's kappa statistic was not statistically significant and most models were still not reliable after pruning. According to Geurts (2000), pruning reduces the complexity of the trees significantly and the variance to some extent. On the other hand, pruning also increases the bias and thus is able to improve only slightly the accuracy of a model. This could be a possible explanation for the stability of Cohen's kappa statistic after pruning and the relatively low increase of the accuracy of the model. Another explanation can be the small size of the dataset. Additionally, the fact that Cohen's kappa does not improve after pruning could be a strong indication for the inefficiency of the selected input variables to predict the habitat suitability of the macroinvertebrates. In the predictive model for Heptageniidae tree-pruning resulted in a significant increase of the CCI (%), while Cohen's kappa decreased significantly and the model failed to construct a tree. The fact that Heptageniidae has a very low frequency of occurrence in the dataset is the most probable explanation for this failure. According to Goethals et al. (2001) and Manel et al. (2001), the predictive performance of models based on decision trees is strongly related to the frequency of occurrence of the predicted taxa. Models that predict very frequently or rarely occurring organisms have a very high CCI (%) and a very low Cohen's kappa statistic as the predictions are based on probabilistic guesses, while a failure of obtaining a tree is very probable. The unpruned model for the prediction of Heptageniidae leads to the construction of a tree that is simple and can be easily ecologically interpreted. However, this tree does not represent realistic relationships between the habitat requirements of that taxon and the environmental characteristics, as it relies on a very restricted amount of instances to extract this information. This is also indicated by Cohen's kappa statistic. The fact that the pruned model for the prediction of Heptageniidae did not lead to the construction of a tree was an improvement, as the model became more ''honest'' and showed that the extraction of knowledge about the

ecological requirements of the predicted taxon was not possible. The improvement of the transparency of the induced models in all cases demonstrates that tree-pruning can be of advantage when the predictive models are used in decision-making of river restoration and conservation management. Similar conclusions were made by Bratko (1989) and Goethals et al. (2001).

## Bagging and boosting optimisation

According to Geurts (2000), a combination of tree-pruning and bagging may lead to more accurate predictions and even if this does not happen, the computational time is always less, without any decrease of the predictive performance of the model. Therefore, bagging and boosting were applied on the pruned trees for further optimisation of the predictive results.

First, bagging did not seem to have a statistically significant effect on the predictive performance of most of the models. If bagging had a significant effect, it would either increase or decrease the performance. Similarly, the effect of boosting varied in relation to the predicted organism. In the models for the prediction of Asellidae, Baetidae and Gammaridae, the two techniques improved the predictive results, while in the predictive models for Caenidae and Gomphidae the application of the two techniques resulted in a decreased predictive performance. In the case of Heptageniidae, the CCI (%) decreased but the model was able to make more reliable predictions, as indicated by Cohen's kappa statistic. It seems that the effect of bagging and boosting on the models depends strongly on the dataset. The fact that the gain in the predictive performance of models when applying boosting decreases as the noise in the data increases, in addition to the fact that boosting can quickly overfit a dataset are the most probable explanations for the cases that boosting does not improve the predictive performance of some of the models (Maclin and Optiz 1997). Boosting can over-emphasise examples of the dataset that are noisy. In this way, it can reduce the predictive performance. However, there are cases where boosting outperforms the individual classifier and also bagging. According to Maclin and Optiz (1997), bagging is more resistant to the noise in the dataset. In most cases, bagging gives better predictive results than the individual classifier. However, there are also cases where it was not effective. Similar conclusions were also reached by Quinlan (1996).

As it was not possible to make any conclusion about the general effect of bagging and boosting on the models when looking at CCI (%) and Cohen's kappa statistic, the percentage of correctly predicted presence and absence cases with each model was examined. Although there were cases where the two techniques had no effect on the CCI (%), there was a tendency to increase the percentage of correct predictions of absence and decrease the percentage of correct predictions of presence. This function of bagging and boosting could be useful for decision-making in river management if it would increase the percentage of correct predictions of presence. In that case, more ecological information on the habitat requirements of the predicted taxa could be obtained. However, the variability of the effect of the two techniques on different models, implies that their convenience is strongly related to the dataset on which these techniques are applied and this should be examined first. Looking at the effect of the two techniques in different repetitions of 10-fold cross-validations within the same model, an even larger variability was observed. This indicates that the replications of examples in the dataset that are produced by bagging and boosting can even lead to contradictive predictions, which make the function of the two techniques on this dataset less clear. The possible outcomes of the application of bagging and boosting should be carefully examined, as in some cases it improves and in some others it reduces the accuracy of the models, as well as their practical applicability. Also, the fact that both techniques produced complicated results, which could not be easily ecologically interpreted, reduced their suitability for management purposes even more.

## Practical applications

Regarding practical applications of classification trees in water management, the set of studies

related to macroinvertebrates is very limited. Practical studies were established by D'heygere et al. (2002) and Goethals et al. (2002). Both studies can only called 'preliminary', seen the small datasets that were available for the studies. D'heygere et al. (2002) researched the use of classification trees to set up a monitoring network in the Dender river (Flanders, Belgium) for the implementation of the European Water Framework Directive (EU 2000). In particular the effect of seasonality was analysed. In this manner, the trees could help to reduce the sampling costs, seen not for all stream types, a multi-seasonal sampling seemed to be interesting due to the very poor ecology present in the Dender. The study of Goethals et al. (2002) aimed at analysing the ecological niches of macroinvertebrates in the Zwalm river basin (Flanders, Belgium) and check the convenience of these models to make predictions on river restoration projects. Classification trees were constructed for all taxa collected during the 60 samplings in the headwaters of the Zwalm river basin. The poor performance of most induced trees had probably its origin in the small size of the dataset. Therefore, also other methods will be applied (ANN, support vector machines and multivariate statistics), that can maybe better deal with the size of the dataset or reveal other specificities of the collected data.

This study therefore illustrated that three types of model validation are at least necessary to make sure that this type of models can be used in water management: theoretical validation based on well chosen performance indicators (thus also taking care of the prevalence of the taxa), comparison with existing ecological knowledge and practical simulation exercises. On top of this, the willingness of river managers to use these data driven ecological models for practical applications is rather low, because of lack of transparency and difficulties with predictions outside the training range.

## Conclusions

This paper explores the induction of models for the prediction of the habitat suitability of six benthic macroinvertebrate taxa in the Axion river in Northern Greece. An evaluation is made of three model optimisation techniques, namely tree-pruning, bagging and boosting. The results of the present study demonstrate that although the models intrinsically proved to have a relatively high predictive power, the noise in the dataset and the inappropriate input variables actually prevented them to some extend, from making reliable predictions. Although tree-pruning did not seem to improve significantly the reliability of the induced models, it reduced considerably the tree complexity. In this way, it increased the transparency of the trees, allowing for a clear ecological interpretation of the induced models. It can thus be concluded that tree-pruning has a high potential when applied to models used for decision-making in river restoration and conservation management. The effect of bagging and boosting varied considerably between the different models, as well as within different repetitions of 10-fold cross-validation in an individual model. In some cases the predictive performance was improved, while in others it was stable or even reduced. The effect of bagging and boosting seemed to be strongly dependent on the dataset on which the two techniques are applied. The present study demonstrated that bagging and boosting is capable of improving the predictive performance of ecological models when properly applied, while an application on inappropriate datasets can result in worse predictive performance.

## References

Adriaenssens V, Goethals PLM, Charles J, De Pauw N (2004) Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers. Ann Limnol – Int J Lim 40(3):181–191

Argiropoulos D (1991) Axios River Basin, water quality management, report for the Ministry of Housing, Physical Planning and Environment of Greece. Delft hydraulics, Athens, Greece

Armitage PD, Moss D, Wright JF, Furse MT (1983) The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Res 17(3):333–347

Barros LC, Bassanezi RC, Tonelli PA (2000) Fuzzy modelling in population dynamics. Ecol Model 128:27–33

Bratko I (1989) Machine learning. In: Gilhooly KJ (ed) Human and machine problem solving. Pelnum Press, New York and London, pp 265–287

Breiman L (1996) Heuristics of instability and stabilisation in model selection. Ann Stat 24:2350–2383

Breimann L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Pacific Grove, Wadsworth

Chatzinikolaou G (2001) Monitoring the ecological quality of the waters and the habitat structure of the Axios River. M.Sc. thesis, Aristotle University of Thessaloniki, Thessaloniki, Greece (in Greek)

Chatzinikolaou G (2002) Monitoring the quality of the waters and the quality of habitats of the Axios River in Greece and FYROM. Final report of the Project DAC: Transboundary co-operation and actions for the protection and management of the waters of the Axios River, by order of the Ministry of Housing, Physical Planning and Environment of Greece. Laboratory of Zoology, School of Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece (in Greek)

Clark P, Niblett T (1989) The CN2 induction algorithm. Machine Learn 3(4):261–283

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

D'heygere T, Goethals PLM, De Pauw N (2002) Optimisation of the monitoring strategy of macroinvertebrate communities in the river Dender, in relation to the EU Water Framework Directive. Sci World J 2:607–617

D'heygere T, Goethals PLM, De Pauw N (2003) Genetic algorithms for optimisation of predictive ecosystem models based on decision trees and neural networks. Ecol Model 160:291–300

Dzeroski S, Drumm D (2003) Using regression trees to identify the habitat preference of sea cucumber (*Holothuria leucospilota*) on Rarotonga, Cook Island. Ecol Model 170:219–226

Dzeroski S, Grbovic J, Walley WJ (1997) Machine learning applications in biological classification of river water quality. In: Michalski RS, Bratko I, Kubat M (eds) Machine learning and data mining: methods and applications. John Wiley and Sons Ltd., New York, USA, pp 429–448

Dzeroski S, Demsar D, Grbovic J (2000) Predicting chemical parameters of river water quality from bioindicator data. Appl Intell 13(1):7–17

EU (2000) Directive of the European Parliament and of the Council 2000/60/EC establishing a framework for community action in the field of water policy, Rep. No. PE-CONS 3639/1/00 REV 1. European Union, Luxembourg

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49

Geurts P (2000) Some enhancements of decision tree bagging. In: Zighed DA, Komorowski J, Zytkow J (eds) Proceedings of the 4th European conference on principles of data mining and knowledge discovery. Springer-Verlag, Berlin, pp 136–147

Goethals PLM (2005) Data driven development of predictive ecological models for benthic macroinvertebrates in rivers. PhD thesis, Faculty of Biosciences Engineering, Ghent University, Gent, 400 pp

Goethals P, De Pauw N (2001) Development of a concept for integrated ecological assessment in Flanders, Belgium. J Limnol 60:7–16

Goethals PLM, Džeroski S, Vanrolleghem P, De Pauw N (2001) Prediction of benthic macro-invertebrate taxa (Asellidae and Tubificidae) in watercourses of Flanders by means of classification trees. In: IWA 2nd World water congress, Berlin, pp 5–6

Goethals PLM, Dedecker AP, Gabriels W, De Pauw N (2002) Development and application of predictive river ecosystem models based on classification trees and artificial neural networks. In: Recknagel F (ed) Ecological informatics: understanding ecology by biologically-inspired computation. Springer-Verlag, Berlin, 432 pp

Kampa E, Artemiadou V, Lazaridou-Dimitriadou M (2000) Ecological quality of river Axios (N.Greece) during spring and summer, 1997. Belg J Zool 130:23–29

Kohavi R (1995) A study of cross-validation and bootstrap for estimation and model selection. In: Mellish CS (ed) Proceedings of the 14th international joint conference on artificial intelligence. Morgan Kaufmann Publishers, Montreal, pp 1137–1143

Kompare B, Bratko I, Steinman F, Dzeroski S (1994) Using machine learning techniques in the construction of models. Part I: introduction. Ecol Model 75–76:617–628

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Langrick JM, Artemiadou V, Yfantis G, Lazaridou-Dimitriadou M, White KN (1998) An integrated water quality assessment of the river Axios, Northern Greece. In: Proceedings of the international conference ''Protection and restoration of the environment IV'', Halkidiki, vol. 1. pp 135–143

Lek S, Guegan JF (1999) Artificial neural networks as a tool in ecological modelling, an introduction. Ecol Model 120:65–73

Maclin R, Optiz D (1997) An empirical evaluation of bagging and boosting. In: Proceedings of the 14th American association for artificial intelligence national conference on artificial intelligence. AAAI Press, California, pp 546–551

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931

Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Francisco, 302 pp

Quinlan JR (1996) Bagging, boosting and C4.5. In: Proceedings of the 13th American association for artificial intelligence national conference on artificial intelligence. AAAI Press, California, pp 725–730

Recknagel F (2002) Ecological informatics: understanding ecology by biologically-inspired computation. Springer-Verlag, Berlin, 432 pp

Witten IH, Frank E (2000) Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers, San Francisco, 371 pp