Ecological Informatics 5 (2010) 147–152

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Comparison of modelling techniques to predict macroinvertebrate community composition in rivers of Ethiopia

Argaw Ambelu, Koen Lock *, Peter Goethals

Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, Belgium

ARTICLE INFO

Keywords:
Classification tree
Support vector machines
Model optimisation
Greedy stepwise
Genetic algorithm

ABSTRACT

In order to fulfil the millennium development goals and to ensure environmental sustainability in Ethiopia, ecological indicator systems can support river managers to analyse the status of watercourses and to select critical restoration actions. In order to use macroinvertebrates as river water quality monitoring and assessment tools, Ethiopia needs data from reference as well as disturbed conditions of surface water ecosystems. Macroinvertebrates, structural and physical–chemical data were in this context collected in the Gilgel Gibe river basin in South-Western Ethiopia during the period 2005–2008. In the next stage, ecological metrics were compared for their assessment relevance. In the present paper, classification trees and support vector machines were used to induce models describing the relation between the river characteristics and the ecological conditions of these streams. Greedy stepwise and genetic search algorithms improved the performance and easy interpretation of these models by making a selection of the variables that were used as input of these models. The developed models allowed to identify the major variables affecting river quality. These tools can support river managers in their decision-making regarding the status of rivers and potential restoration options, for example by providing rules concerning critical values of major river characteristics at which certain actions should be undertaken.

1. Introduction

In recent years, predictive models have been applied in numerous ecological studies to assess, monitor and manage natural resources by considering the presence or absence of indicator organisms in a given habitat (Goethals et al., 2002; D'heygere et al., 2003, 2006; Dedecker et al., 2005; Dakou et al., 2007). Predictive modelling is one of the most important steps in the development of a standard habitat assessment protocol (Parsons et al., 2004). Classification trees (CT) and support vector machines (SVM) are two examples of predictive modelling techniques that can be used in this context.

Due to their transparency and flexibility, classification trees are recently gaining popularity. Classification trees are attractive since they are fairly straightforward to construct and their transparency allows for easy integration into an environmental decision support system. In environmental applications, classification trees have been used to predict biological indicators (Edwards et al., 2006). Dzeroski et al. (1997) were among the first to describe applications of classification trees in river community analysis. Since then, they have also been applied to predict macroinvertebrate presence (D'heygere et al., 2003; Dakou et al., 2007). Such predictions may provide a good indication of the impact of different anthropogenic disturbances on river conditions (Ghetti and Ravera, 1993; Rosenberg and Resh, 1993; Hall and Holmes, 2003; D'heygere et al., 2006; Dakou et al., 2007; Goethals et al., 2007).

SVM are a new generation of learning algorithms that implement Platt's sequential minimal optimisation (SMO) algorithm for training a support vector classifier (Keerthi et al., 2001). They deal with many predictors and they also avoid the assumption of linear relationships (Akkermans et al., 2005). SVM are a group of supervised machine learning methods that can be applied for classification or regression algorithms. They replace all missing values and transform nominal attributes into binary ones (Witten and Frank, 2000). Due to this reason, SVM are claimed to have a good model performance compared to other techniques. For example, Akkermans et al. (2005) compared the predictive performance of SVM and logistic regressions to predict macro-fauna community types from environmental variables and SVM were found to be the best predictor. This made SVM an interesting modelling tool (Vapnik, 1995; Burges, 1998; Keerthi et al., 2001) because of the very good performance in different fields of application (Dibike et al., 2001; Keerthi et al., 2001).

Irrelevant attributes in a dataset are known to reduce model performance and reliability (Hall and Holmes, 2003) and if there are too many attributes, they are costly and are not manageable (Dom et al., 1989). Greedy stepwise and genetic algorithms are some of the known search algorithms to select input attributes (Gevrey et al., 2003; Gabriels et al., 2007). The greedy stepwise or hill-climbing

* Corresponding author. J. Plateaustraat 22, 9000 Ghent, Belgium. Fax: +32 9 2643766.
E-mail address: Koen.Lock@UGent.be (K. Lock).

1574-9541/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2009.12.004

attribute selection approach considers both adding and removing features at each decision point, which allows to retract an earlier decision without keeping explicit track of the search path (Blum and Langley, 1997). After features are generated, one can select the best or simply, the first feature that improves accuracy over the current set (Blum and Langley, 1997; Boros et al., 2003). Greedy algorithms have been used to select the appropriate attributes and also to remove variables (Butterworth et al., 2004). Genetic algorithms are general purpose search algorithms inspired by Charles Darwin's principle of 'survival of the fittest' to solve complex optimisation problems (Holland, 1975; Goldberg, 1989; Vose, 1999). They got popular for the optimisation of predictive models, specifically in the field of river ecology (D'heygere et al., 2006; Goethals et al., 2007).

River management can cause a complex decision-making process because it has to consider the natural physical system of the river catchment, the socioeconomic system that relies upon the water resources of a given river basin, the administrative body responsible for river management etc. These complex decision processes could be simplified by developing different decision support systems (Welp, 2001). Predictive models for macroinvertebrate taxa or metrics that are sensitive to specific river basin problems could be useful decision support tools. CT and SVM are some of the modelling methods that can be used to predict macroinvertebrate compositions in river systems. From such models, relevant parameters that should be considered by river managers during river restoration activities could be derived. The present paper aimed to predict the macroinvertebrate metrics best reflecting the ecological water quality in South-Western Ethiopia using SVM and CT.

2. Materials and methods

2.1. Study area

The Gilgel Gibe watershed is located in south-western Ethiopia (latitude 7°25′–7°55′ North and longitude 36°30′–37°22′ East), while the altitude ranged from 1096 to 3259 m. The area is mainly used for agricultural and other anthropogenic activities. Almost the whole watershed is exposed to different activities like grazing, ploughing, sand dredging, vegetation clearance and municipal waste discharge. In general, most watercourses are exposed to point and non-point pollution sources. However, there were some stream segments which are still covered with a natural vegetation and remote from direct human activity. The location of the sampling sites in the Gilgel Gibe watershed in Ethiopia is illustrated in Fig. 1.

2.2. Data collection

Macroinvertebrates and environmental data were collected at 42 sampling sites in streams of the Gilgel Gibe watershed from January 2005 till September 2008. Depending on the sampling site, one to seven samples were taken. During the whole campaign, 162 samples were collected with the kick sampling method as described by Gabriels et al. (2009) using a D-frame net having a mesh size of 300 µm diameter. During 5 min, samples were collected from each habitat within a 10 m stretch. Macroinvertebrates were then sorted alive onsite and preserved in 70% ethanol for identification.

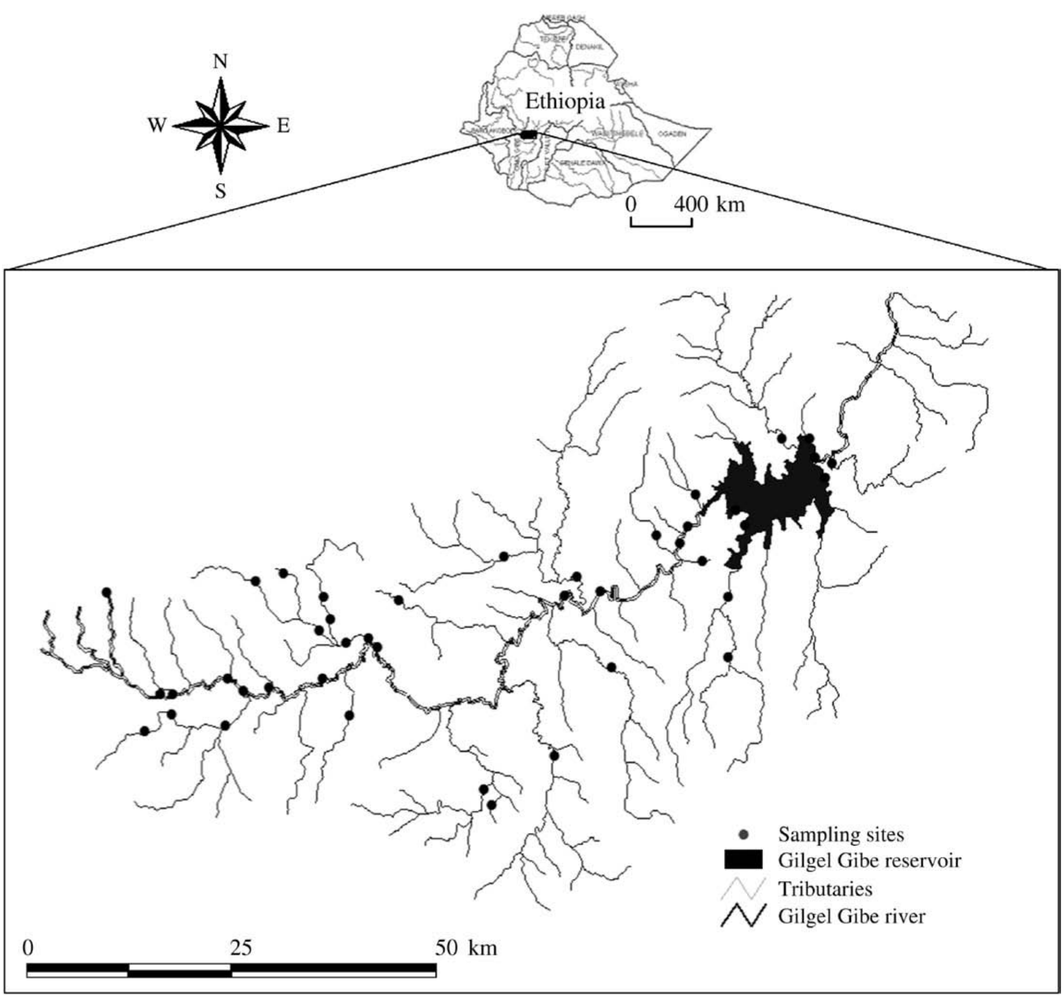

Fig. 1. Location of the sampling sites in the Gilgel Gibe watershed in Ethiopia.

To assess the river habitat status of the sampling reach, a survey was undertaken with the intention of identifying those physical features of the river that have direct or indirect influence on the macroinvertebrate community. The habitat of each sampling reach was characterized using the USEPA rapid physical habitat classification format (Barbour et al., 1999). Habitat scores of the sampling sites were used to classify the site as poor (<60), marginal (60–99), suboptimal (100–159) and optimal (160–200). Using Statistica (Statsoft®), a multivariate discriminant analysis was performed based on the log transformed physical–chemical variables and using a forward stepwise selection method together with canonical analysis. The discriminant analysis allowed to check the appropriateness of the application of the USEPA rapid physical habitat classification format in the studied watershed.

Physical–chemical parameters like temperature and conductivity (Knick Portamess® 911 conductivity meter), pH (Knick Portamess® 911 pH meter), oxygen saturation (Knick Portamess® 911 oxygen probe) and turbidity (portable Wagtech® turbidity meter) were measured onsite at each sampling location. Five day biochemical oxygen demand (BOD_5), nitrate-nitrogen (described as nitrate), ammonium-nitrogen (described as ammonium) and orthophosphate (described as phosphate) were measured in the laboratory according to standard methods (APHA, AWWA, WPCF, 1995). The distance between the sampling site and the source was measured using GIS and altitude was measured using the Global Positioning System (Magellan®, SporTrak Pro). In total, 22 different river characteristics were recorded (Table 1).

2.3. Metric classification

As it is difficult to have a general model for all macroinvertebrates, the use of metrics that best suit the regional application is mandatory (Wagner et al., 2006). Among many macroinvertebrate metrics, the % organisms belonging to Ephemeroptera, Plecoptera and Trichoptera (% EPT), the % organisms being scrapers, the Biological Monitoring Working Party score (BMWP) and taxa richness (TR) were chosen because these metrics are easy to apply and are therefore suitable for developing countries (Resh, 2007) and because they best reflected the ecological water quality (Ambelu, 2009). Four classes were distinguished for each macroinvertebrate metric (Table 2, Fig. 2).

Table 1
Input variables used for the model development together with the average, standard deviation, minimum and maximum values.

Input variables	Unit	Average	St. dev.	Minimum	Maximum
Altitude	m	1737	87	1625	2488
Stream order		2.4	1.1	1.0	4.0
Riparian vegetation status	Score (0–20)	7	6	0	20
Distance from source	km	32	36	2	154
Sinuosity	score (0–20)	14	4	7	20
Velocity	m/s	0.50	0.36	0.01	1.80
Discharge	m^3/s	2.01	2.21	0.001	12.32
Average water depth	m	0.60	0.41	0.01	2.00
Wetted width	m	9	9	1	43
Water temperature	°C	20.5	2.5	15.2	28.5
Ambient temperature	°C	24	3	16	32
Embeddedness	Score (0–20)	12	6	0	19
Riverbank status	Score (0–20)	12	7	1	26
Turbidity	FTU[a]	80	135	0	1000
pH		7.3	0.5	5.3	8.5
Conductivity	µS/cm	113	54	40	450
Dissolved oxygen	mg/l	6.4	1.5	1	9.3
Oxygen saturation	%	85	23	7	132
BOD	mg/l	3.5	4.6	0.5	26.0
Ammonium	mg/l	0.80	0.65	0.01	2.60
Nitrate	mg/l	1.50	1.08	0.01	4.80
Phosphate	mg/l	0.45	0.50	0.00	2.17
% Ephemeroptera, Plecoptera and Trichoptera		Four class (poor, fair, good and high)			
% scrapers		Four class (poor, fair, good and high)			
Biological Monitoring Working Party score		Four class (poor, fair, good and high)			
Taxa richness		Four class (poor, fair, good and high)			

[a] Formazine turbidity unit.

2.4. Model training and validation

SVM and CT models were developed using Weka Platt's sequential minimal optimisation (SMO) and the J48 algorithm (Witten and Frank, 2000), respectively. For the training and validation of SVM and CT models, 10-fold cross-validation was used to get a reliable estimate of the error of each model (Kohavi, 1995; Witten and Frank, 2000; Dakou et al., 2007). This procedure is helpful to avoid overfitting of the models (Bishop 1995). To evaluate the performance of each method, percentage of correctly classified instances (% CCI) and kappa statistics (K) were used (Witten and Frank, 2000; Goethals et al., 2007).

2.5. Attribute selection and optimisation

Selection of the appropriate variables in a dataset is important because it enhances model performance (Goethals et al., 2007). In this study, optimisation of the SVM and CT models were made by applying greedy and genetic algorithms. Initially, models were run using all 22 input variables. Subsequently, variables were selected by greedy stepwise and genetic algorithms for the four selected macroinvertebrate metrics. Wrapper subset evaluator using SMO and J48 as a base learning algorithm were used. The wrapper subset evaluator evaluates variables by using accuracy estimates provided by the actual target learning algorithm (Hall and Holmes, 2003). During model building, default settings were used of the Weka software package in Java (Witten and Frank, 2000). After randomization, each model was run five times for each macroinvertebrate metric in order to check the consistency and reproducibility of the model. The predictive performance of each technique for the four macroinvertebrate metrics was calculated from the output of 10-fold cross-validation. The Kruskal–Wallis tests were applied to assess whether there were differences in performance between the different model optimisation techniques.

3. Results

Multivariate discriminant canonical analysis based on the physical–chemical variables showed that the four habitat groups were distinctly clustered (Fig. 3). The rapid habitat assessment protocol thus proved to be an appropriate tool in the context of this study. The cumulative proportion of the first and the second canonical roots represented 86% and 96% of all eigenvalues, respectively. The canonical R values of the first and second roots were respectively 0.94 and 0.70, which indicates a good representation of the underlying resemblance, and these two roots significantly discriminated between the four habitat categories ($p<0.00001$).

Table 2
Division in four classes of the used macroinvertebrate metrics: % organisms belonging to Ephemeroptera, Plecoptera and Trichoptera (EPT), % organisms being scrapers, Biological Monitoring Working Party score (BMWP) and taxa richness (TR).

	Poor	Fair	Good	High
% EPT	<10	10–40	>40–60	>60
% scrapers	<6	6–20	>20–40	>40
BMWP	<30	30–50	51–100	>100
TR	<7	7–12	13–25	>25

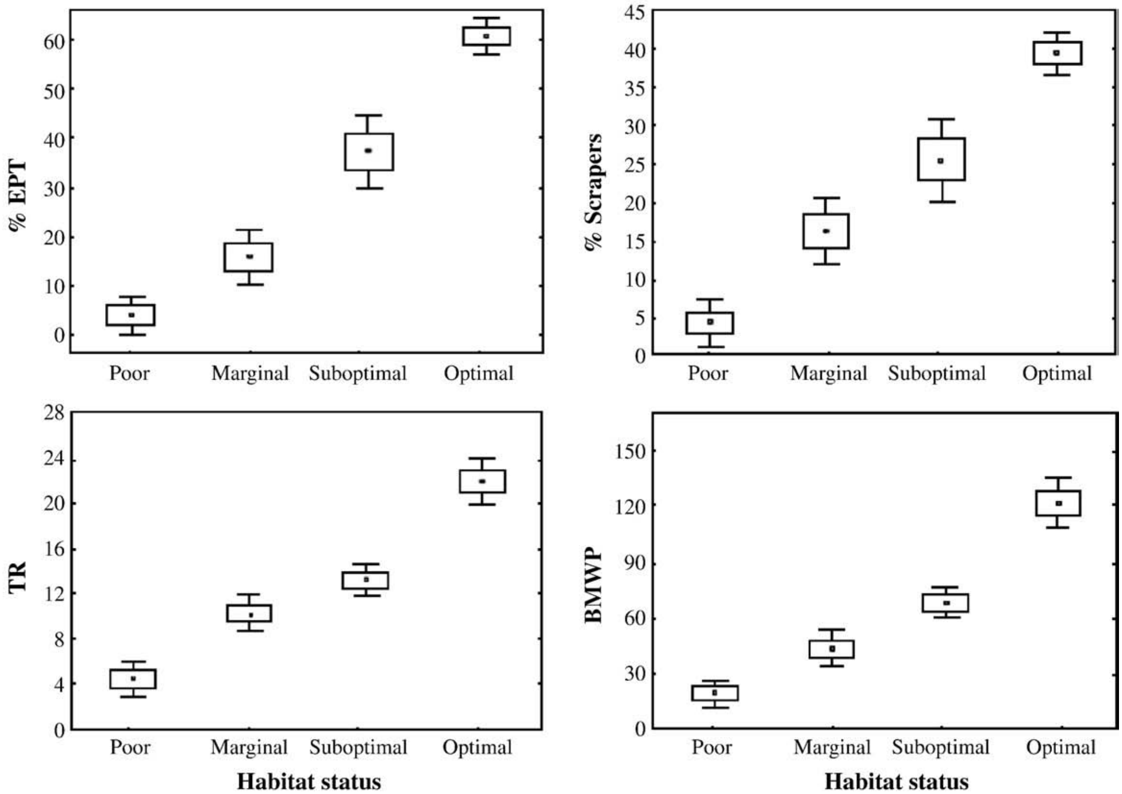

Fig. 2. Box (standard error) and whisker (95% confidence interval) plot of the predicted macroinvertebrate metrics (% organisms belonging to Ephemeroptera, Plecoptera and Trichoptera (EPT), % organisms being scrapers, Biological Monitoring Working Party score (BMWP) and taxa richness (TR)) for the four recognized habitat quality classes.

The performance of SVM and CT models was compared in three ways: (1) between SVM and TC models, (2) before and after the application of optimisation methods and (3) between the different macroinvertebrate metrics. Before optimisation, using all the input variables, both SVM and CT models performed less good than after variable selection (Table 3). However, when the two techniques were compared, SVM showed significantly better performance over CT ($p<0.001$). Significant model improvements ($p<0.05$) were also observed with both techniques after the application of greedy and genetic search algorithms in terms of both % CCI and *K*. After optimisation, performance of SVM and CT models were still different, but significant differences were no longer observed ($p>0.05$). Among the four macroinvertebrate metrics, % scrapers was poorly predicted and its kappa statistics were less than 0.4 while for the other metrics, it was higher than 0.4 (Table 3).

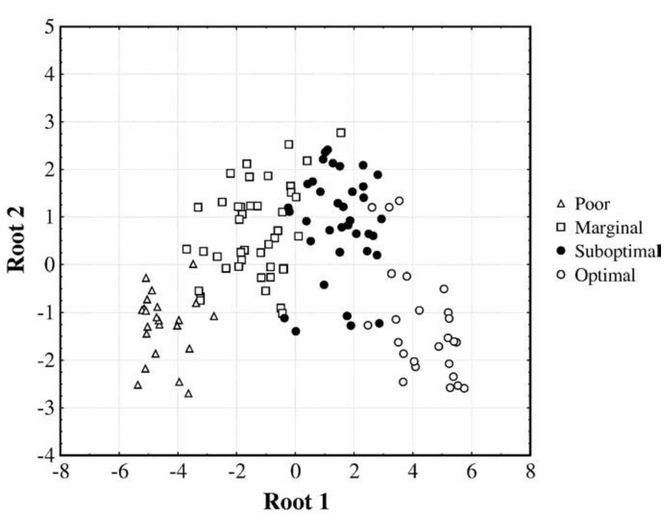

Fig. 3. Scatter plot of canonical discriminant analysis based on the environmental variables used for modelling, with indication of the four distinguished habitat categories.

The performance of the six modelling methods (SVM and CT, both without optimisation, with greedy stepwise and with genetic algorithms) were compared by looking at the model performance indicators (CCI and *K*) of the four metrics together (Table 3). Without the application of optimisation techniques, the SVM performed better than the CT. After greedy stepwise optimisation, SVM again showed superior performance over CT, however, after application of genetic algorithms, SVM and CT showed a similar performance in terms of *K*. SVM showed smaller standard deviations for % CCI and *K* in comparison with CT (Table 3), indicating that SVM gave more stable results. When both optimisation methods were compared, the greedy stepwise search algorithm performed better for SVM, while for CT, the genetic search algorithm resulted in better predictive models. The application of search algorithms made CT more transparent because limiting the number of input variables resulted in a reduction of the size of the induced trees. In the CT model for % EPT, for example, both greedy and genetic algorithms reduced the size of the tree from 55 to 21 and the number of leaves from 28 to 11.

The Kruskal–Wallis rank test revealed that there was a significant model performance difference between models made with optimisation and without optimisation techniques. The two search algorithms selected the most important variables predicting the macroinvertebrate metrics. Structural variables of the river, such as embeddedness, riverbank status and the presence of riparian vegetation, as well as physical–chemical parameters, such as pH and water temperature, were selected by both methods (Table 4). In general, the number of attributes retained by genetic search algorithms was relatively higher than by greedy stepwise algorithms.

Table 3
Average ($n = 5$) model performance for the used metrics (% organisms belonging to Ephemeroptera, Plecoptera and Trichoptera (EPT), % organisms being scrapers, Biological Monitoring Working Party score (BMWP) and taxa richness (TR)) for the support vector machines and classification trees; No = no optimisation method applied, Gr = greedy stepwise and Gen = genetic algorithm. The mean value indicates the average performance for the four metrics (standard deviation between brackets).

	Support vector machines						Classification trees					
	No		Gr		Gen		No		Gr		Gen	
	% CCI	*K*	% CCI	*K*	% CCI	*K*	% CCI	*K*	% CCI	*K*	% CCI	*K*
% EPT	56	0.39	59	0.43	57	0.40	50	0.31	59	0.43	60	0.44
% scrapers	48	0.27	54	0.36	54	0.36	42	0.22	47	0.26	51	0.32
BMWP	57	0.39	61	0.43	61	0.44	53	0.35	60	0.44	64	0.50
TR	58	0.37	65	0.48	63	0.46	51	0.30	60	0.40	59	0.42
Mean	55 (4.6)	0.36 (0.057)	60 (4.6)	0.43 (0.049)	59 (4.0)	0.42 (0.044)	49 (4.8)	0.30 (0.054)	57 (6.4)	0.38 (0.083)	59 (5.4)	0.42 (0.075)

4. Discussion

The composition of the benthic macroinvertebrate communities usually provides useful insights in the ecological quality of surface waters, as these organisms are sensitive to disturbance. Predicting the composition of macroinvertebrate communities in rivers is not an easy task, because of the number of species that can be modelled and due to the complexity of biotic and abiotic variables that influence their distribution. However, predicting macroinvertebrate metrics like % EPT, % scrapers, BMWP and taxa richness, which are metrics that are known to be well correlated with the ecological water quality, can result in comprehensive models which are helpful for decision support systems.

In the present study, the prediction of these macroinvertebrate metrics by SVM and CT resulted in reliable models. Before optimisation, SVM performed better than CT models. This was probably due to the robustness of SVM over CT: SVM are less affected by missing data and multiple collinearity (Witten and Frank, 2000). The performance of CT was somewhat lower (below 51% CCI and $K = 0.32$) and in addition, CT were complex and therefore difficult to interpret. In most data mining techniques, such as artificial neural networks and CT, the model performance can be compromised by collinearity and noisy datasets (D'heygere et al., 2006; Goethals et al., 2007). The quality of models developed by data mining techniques can also be compromised if the dataset contains too much irrelevant or unreliable information (Hall and Holmes, 2003). After the application of greedy stepwise or genetic search algorithms, models always showed better performance. Models performed best when CT were combined with genetic algorithms and SVM with greedy stepwise algorithms. In the CT models, greedy search algorithm were probably trapped by local noise data so global maxima were not reached (Vafaie and Imam, 1994). The application of the two algorithms not only helped to optimize the models, but also made them easier to interpret. Although the performance of SVM was generally slightly better, CT are more easily applied by non-specialists, which is a great advantage, especially in developing countries.

It should be kept in mind that the variables selected for the model development are not necessarily the only ones that are important. Variables that are not selected can be poor predictors, but another possibility is that they are correlated with another variable or with a combination of other variables. Variable selection during model development thus indicates whether a variable is important, but it is not decisive on its own. In fact, principal component analysis indicated that nutrient concentrations, such as ammonium, phosphate and nitrate, were strongly correlated and also parameters like riparian vegetation cover, quality of the riverbank and embeddedness were strongly related (Ambelu, 2009).

Table 4
Selected variables (marked X) by greedy (Gr) and genetic (Gen) algorithms applied for support vector machines (SVM) and classification trees (CT) for the prediction of the four metrics: % organisms belonging to Ephemeroptera, Plecoptera and Trichoptera (EPT), % organisms being scrapers, Biological Monitoring Working Party score (BMWP) and taxa richness (TR).

	% EPT				% scrapers				BMWP				TR			
	SVM		CT		SVM		CT		SVM		CT		SVM		CT	
	Gr	Gen	Gr	Gen	Gr	Gen	Gr	Gen	Gr	Gen	Gr	Gen	Gr	Gen	Gr	Gen
Altitude	X	X					X	X			X			X		X
Stream order									X	X			X	X	X	
Riparian vegetation		X	X	X	X			X	X	X				X		
Distance from source			X													
Sinuosity					X	X	X									
Velocity	X			X												
Discharge		X		X				X				X				
Water depth		X														
Wetted width									X	X			X	X		X
Water temperature		X		X					X	X	X	X				
Ambient temperature														X		
Embeddedness	X	X	X	X	X	X						X	X	X	X	X
Riverbank status		X					X		X		X	X	X	X	X	X
Turbidity	X									X		X				
pH	X				X	X			X	X					X	X
Conductivity						X										
Dissolved oxygen														X		
Oxygen saturation	X	X								X						X
BOD											X	X				
Ammonium	X															
Nitrate		X	X	X	X					X						
Phosphate	X	X													X	

Dzeroski et al. (1997) modelled the biological classification of British rivers based on biological data, 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. Blockeel et al. (1999) made simultaneous predictions of multiple physical–chemical properties of the water from its biological properties using a single decision tree and also predictions of past physical–chemical properties of the river water from its current biological properties, while Dzeroski et al. (2000) predicted physical–chemical variables on the basis of biological data, taxa that occurred in many trees were considered as useful indicator taxa. These studies therefore use the opposite approach as was conducted in the present study, but with the same objective. D'heygere et al. (2003, 2006) predicted the presence of several macroinvertebrate taxa in Flanders based on a selection of environmental variables, while Dakou et al. (2007) induced decision trees to predict the habitat suitability of six macroinvertebrate taxa in the river Axios (Northern Greece). Although these latter models were able to predict the occurrence of a single species, they are less suitable to judge the general ecological water quality. It can thus be concluded that machine learning techniques such as CT and SVM can be used in various ways to assist in water quality management.

5. Conclusions

In the present study, SVM and CT with and without optimisation methods were used to induce models predicting macroinvertebrate metrics. The knowledge obtained in this way could be helpful for river management and restoration purposes. Based on the developed models, it can be concluded that acting on the restoration of the riparian vegetation and minimisation of nutrient input and organic waste discharges would considerably improve the ecological quality in the Gilgel Gibe watershed in Ethiopia.

Acknowledgments

The authors are grateful to VLIR-UOS for financing this study. The first author is a recipient of ICP-PhD scholarship from this same organization. Koen Lock is currently supported by a post-doctoral fellowship from the Fund for Scientific Research (FWO-Vlaanderen, Belgium).

References

Akkermans, W., Verdonschot, P., Nijboer, R., Goedhart, P., ter Braak, C., 2005. Predicting macro-fauna community types from environmental variables by means of support vector machines. In: Lek, S., Scardi, M., Verdonschot, P., Descy, J., Park, Y.S. (Eds.), Modelling Community Structure in Freshwater Ecosystems. Springer, Berlin, pp. 411–434.
Ambelu, A.B., 2009. Biological Monitoring Based on Macroinverteabrates for Decision Support of Water Management in Ethiopia. Ghent University, Ghent.
APHA, AWWA, WPCF, 1995. Standard Methods for the Examination of Water and Wastewater, 19th ed. American Public Health Association, Washington D.C.
Barbour, M.T., Gerritsen, J., Snyder, B.D., Stribling, J.B., 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macro-invertebrates and Fish, Second Edition. U.S. Environmental Protection Agency, Office of Water, Washington D.C.. EPA 841-B-99-002.
Bishop, C., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.
Blockeel, H., Dzeroski, S., Grbovic, J., 1999. Simultaneous prediction of multiple chemical parameters of river water quality with TILDE. Lecture Notes in Artificial Intelligence 1704, 32–40.
Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271.
Boros, E., Horiyama, T., Ibaraki, T., Makino, K., Yagiura, M., 2003. Finding essential attributes from binary data. Annals of Mathematics and Artificial Intelligence 39, 223–257.
Burges, C.J.C., 1998. Geometry and invariance in kernel based methods. In: Scholkopf, B., Burges, C.J.C., Smola, A.J. (Eds.), Advances in Kernel Methods — Support Vector Learning. MIT Press, Cambridge.
Butterworth, R., Simovici, D.A., Santos, G.S., Ohno-Machado, L., 2004. A greedy algorithm for supervised discretization. Journal of Biomedical Informatics 37, 285–292.
Dakou, E., D'heygere, T., Dedecker, A.P., Goethals, P.L.M., Dimitriadou, M.L., De Pauw, N., 2007. Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). Aquatic Ecology 41, 399–411.
Dedecker, A.P., Goethals, P.L.M., D'heygere, T., Gevrey, M., Lek, S., De Pauw, N., 2005. Application of artificial neural network models to analyse the relationships between *Gammarus pulex* L. (Crustacea, Amphipoda) and river characteristics. Environmental Modelling and Assessment 111, 223–241.
D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Genetic algorithms for optimisation of predicitive ecosystem models based on decision trees and neural networks. Ecological Modelling 160, 291–300.
D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecological Modelling 195, 20–29.
Dibike, Y.B., Velickov, S., Solomatine, D., Abbott, M.B., 2001. Model induction with support vector machines: introduction and applications. Journal of Computing in Civil Engineering 15, 208–216.
Dom, B., Niblack, W., Sheinvald, J., 1989. Feature selection with stochastic complexity. Proceedings of IEEE on Computer Vision and Pattern Recognition, Rosemont, pp. 241–248.
Dzeroski, S., Grbovic, J., Walley, W.J., 1997. Machine learning applications in biological classification of river water quality. In: Michalski, R.S., Bratko, I., Kubat, M. (Eds.), Machine Learning and Data Mining: Methods and Applications. John Wiley and Sons Ltd., New York, pp. 429–448.
Dzeroski, S., Demsar, D., Grbovic, J., 2000. Predicting chemical parameters of river water quality from bioindicator data. Applied Intelligence 13, 7–17.
Edwards, T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Moisen, G.G., 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecological Modelling 199, 132–141.
Gabriels, W., Goethals, P.L.M., Dedecker, A.P., Lek, S., De Pauw, N., 2007. Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networds. Aquatic Ecology 41, 427–441.
Gabriels, W., Lock, K., Goethals, P.L.M., 2009. Multimetric Macroinvertebrate Index Flanders (MMIF) for Biological Assessment of Rivers and Lakes in Flanders (Belgium). Limnologica 39.
Gevrey, M., Dimopoulos, L., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological Modelling 160, 249–264.
Ghetti, F.P., Ravera, O., 1993. A European perspective on biological monitoring. In: Loeb, S., Spacie, A. (Eds.), Biological Monitoring of Aquatic Systems. Lewis Publishers, Boca Raton.
Goethals, P.L.M., Dedecker, A.P., 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, pp. 203–213.
Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology 41, 491–508.
Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, Reading.
Hall, M.A., Holmes, G., 2003. Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering 15, 1437–1447.
Holland, J.H., 1975. Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press, Ann Arbor.
Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K., 2001. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation 13, 637–649.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Lavrac, M., Wrobel, S. (Eds.), Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1137–1143.
Parsons, M., Thoms, M.C., Horris, R.H., 2004. Development of a standard approach to river habitat assessment in Australia. Environmental Monitoring and Assessment 98, 109–130.
Resh, V.H., 2007. Multinational, freshwater biomonitoring programs in the developing world: lessons learned from African and Southeast Asian river surveys. Environmental Management 39, 737–748.
Rosenberg, D.M., Resh, V.H., 1993. Freshwater Biomonitoring and Benthic Invertebrates. Chapman & Hall, New York.
Vafaie, H., Imam, I.F., 1994. Feature selection methods: genetic algorithms vs. greedy-like search. Proceedings of the International Conference on Fuzzy and Intelligent Control Systems, Louisville.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York.
Vose, M.D., 1999. The Simple Genetic Algorithm: Foundations and Theory. The MIT Press, Cambridge.
Wagner, R., Obach, M., Werner, H., Schmidt, H.H., 2006. Artificial neural nets and abundance prediction of aquatic insects in small streams. Ecological informatics 1, 423–430.
Welp, M., 2001. The use of decision support tools in participatory river basin management. Physics and chemistry of the earth (B) Hydrology, Oceans and Atmosphere 26, 535–539.
Witten, I.H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, San Francisco.