Caspian J. Env. Sci. 2011, Vol. 9 No.2 pp. 189~198

CJES
Caspian Journal of Environmental Sciences

[Research]

Application of classification trees-J48 to model the presence of roach (*Rutilus rutilus*) in rivers

R. Zarkami

Dept. of Environmental Science, Faculty of Natural Resources, University of Guilan, P.O.Box 1144, Someh Sara, Guilan, Iran.
E-mail: rzarkami2002@yahoo.co.uk

ABSTRACT
In the present study, classification trees (CTs-J48 algorithm) were used to study the occurrence of roach in rivers in Flanders (Belgium). The presence/absence of roach was modelled based on a set of river characteristics. The predictive performance of the CTs models was assessed based on the percentage of Correctly Classified Instances (CCI) and Cohen's kappa statistics. To find the best model performance, a 3-fold cross validation techniques was applied on the dataset. The effect of Pruning Confidence Factors (PCFs) was examined on the reliability and model complexity. Based on the obtained results, the induced model could predict well the presence/absence of roach in the rivers. The highest overall means of two model performances showed that the models were reliable. When analyzing the ecological relevance of CTs, it seemed that the structural-habitat variables were more the main predictors than the water quality ones to predict the occurrence of roach in rivers. In particular, the distance from the source and width contributed more to the prediction of roach while among water quality variables, only electric conductivity was relatively important in this regard.

Keywords: *Ecological modelling, Classification trees (J48), Occurrence, Roach, River basins, Flanders.*

INTRODUCTION
Until now, prediction of organisms by ecological modelling techniques has been an interesting subject for many researchers (Lawton, 1996; Goethals and De Pauw, 2001; Olden and Jackson, 2002; D'heygere *et al.,* 2003; D'heygere *et al.,* 2006). In particular, habitat and spatial distribution of lake and river fish have been studied for a long time (Copp, 1990; Brosse and Lek, 2000). Habitat suitability models aim to relate the presence or abundance of a species at a site to environmental variables that describe their general habitat. Models predicting presence/absence of organisms are of outstanding importance in freshwater ecosystems (Fielding and Bell, 1997). Habitat use and the specific composition of communities are influenced by interactions between animals and their biotic and abiotic environment (Schoener, 1974). In particular, fish habitat is considered as a significant factor (Werner *et al.,* 1977). Models play key roles for the interpretation of results with more sensitivity and better insight, the simulation of the effect(s) of potential management options and supporting decision-making. Particularly predictive models have various important applications for the conservation and management of fish populations (Goethals *et al.,* 2006). These models are urgently required as the modification and loss of aquatic habitat is now recognized as the key factor threatening the conservation of fish populations and communities all over the world (Ricciardi and Rasmussen, 1999). Models dealing with predictive fish-habitat have an important role in prioritizing surveys and monitoring programmes for fish populations (Jackson and Harvey, 1997).

Roach (*Rutilus rutilus*) L. belongs to Cyprinidea family and is widely distributed over Europe. They are the most abundant fish species in many European lakes. Roach has a relatively rapid growth

(Kahl and Radke, 2006). Since they are classified as omnivorous fish species, they are able to feed on a variety of food resources e.g. epiphytes, phytoplankton, macrophytes, zooplankton, zoobenthos, detritus (Brabrand, 1985; Horppila, 1994; Kahl *et al.*, 2001; Kahl and Radke, 2006). Many studies in relation to the habitat use of roach have been carried out lakes (Skov *et al.*, 2002; Schulze *et al.*, 2006; Kahl and Radke, 2006; Sharma and Borgstrøm, 2008) while less have been done in rivers. The aim of the present study was to develop models that could predict the presence of roach in rivers in Flanders (Belgium) using classification trees and to compare the performance of this technique.

MATERIAL AND METHODS

Study area

The study was conducted in the Flanders situated in the northern part of Belgium. 11 main river basins discharge into Flanders but only 8 river basins were monitored in this study: Nete, Leie, BoSche, Gekan, Dijle, Dender, Demer and Izer. In these river basins, more than 200 sites were chosen for investigation over the study period from 1995 to 2004 (Fig. 1). Agriculture, industrial and domestic wastewaters are main sources of pollution loads in most river basins in particular in Dender resulting in the eutrophication process. Due to artificial embankment, dams and weirs, many freshwater fishes and their migration paths decrease. In contrast to running water, the standing waters of Flanders are intensively contaminated.

More than 50 fish species were recorded by Institute for Forestry and Game Management in rivers in Flanders over the study period from 1995 to 2004. Among these fish species, roach (*Rutilus rutilus*), perch (*Perca fluviatilis* L.), bream (*Abramis brama* (L.)), pike (*Esox lucius* L.), eel (*Anguilla anguilla* (L.)), chub (*Leuciscus cephalus* (L.)), gudgeon (*Gobio gobio* (L.)), rudd (*Scardinius erythrophthalmus* (L.)) and brown trout (*Salmo trutta forma lacustris* L.) were the important ones. Pike, perch and brown trout were the main predatory fish species but this study merely focused on the habitat use of roach that are important prey for pike and perch in the rivers in Flanders.

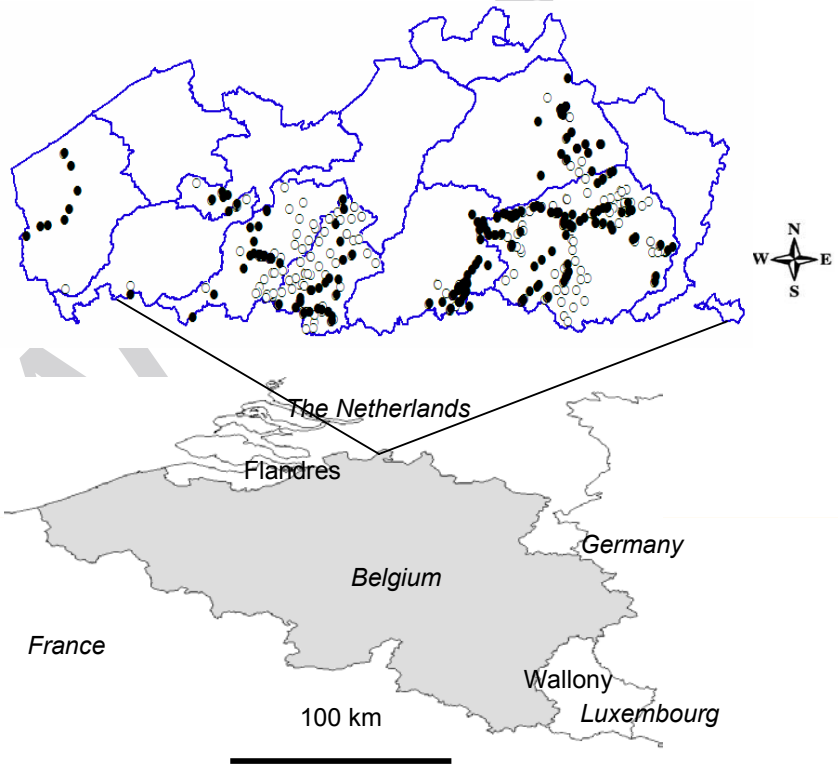

Fig 1. Map of Flanders in Belgium (bottom) and geographical distribution of roach (top) in the sampling sites in the main river basins in Flanders (absence of roach is indicated with solid circle and presence with blank circle) (Goethals, 2005).

Data collection

The sampling points were considered in the main river basins in Flanders. Samples were collected during the day and covered different seasons (on a monthly basis). In total, a large dataset (nearly 600 instances) were available which could serve for the development of classification trees methods to predict the habitat use of roach. A set of river characteristics were recorded such as conductivity, dissolved oxygen, pH, water temperature, gradient slope, river width, river depth, distance from the source, flow velocity, habitat quality, vegetation cover and etc. Some variables were eliminated due to insufficient information such as flow velocity and vegetation cover and habitat quality. And some important water quality variables which represent nutrient content of water courses were also missing such as nitrogen from ammonia ($N\text{-}NH_4^+$), nitrites ($N\text{-}NO_2^-$), nitrates ($N\text{-}NO_3^-$), phosphate ($P\text{-}PO_4^{3-}$) and etc. Distance from the source and slope were measured using overlays in a geographical information system (GIS) and a topographic map. Width and depth were checked and measured in the field. The physical-chemical variables were collectively determined in the field and laboratory on the basis of standardized and quality controlled methods. Table 1 presents summary statistics of some river characteristics that were only used for the model. Biotic variables merely consisted of presence/absence of roach.

Table 1. Summary statistics for the river characteristics used in model development to predict species presence/absence in the study river basins in Flanders (SD: standard deviation).

River characteristics	Unit	Minimum	Maximum	Mean ± SD
Water temperature	˚C	1.40	19.90	11.11 ± 4.23
Distance from the source	Km	0.00	84.80	16.32 ± 18.47
Width	m	0.40	53.30	6.46 ± 7.36
Slope	%	0.00	15	2.20 ± 2.40
Depth	m	0.05	3.10	0.73 ± 0.51
Dissolved oxygen (DO)	mg l^{-1}	0.41	15.50	7.74 ± 2.39
pH		5.2	9.7	7.31 ± 0.57
Electric conductivity (EC)	µS/cm	122	5080	709 ± 384

All the river characteristics were used as input variables while biotic variables could serve as output variables included in the Weka software (version 3-4-11, 1999-2007c; Witten and Frank, 2000). The frequency of occurrence of roach (observed values) was considered 50% in all sites, that is, roach were absent in 50% and also present in 50% of sampling sites. The applied technique to collect fish assemblage data was electrofishing, using a 5 kW generator with an adjustable output voltage of 300 to 500 V and a pulse frequency of 480 Hz. The number of hand-held anodes used was 2 except when the river had a width of less than 1 m. Electrofishing was carried out in 3 Huet zones namely barbel, bream and upstream. This method has been broadly used in ecological fields for fish monitoring and a wide range of river types (e.g. Belpaire *et al.*, 2000; Breine *et al.*, 2004). At each station, roach were counted and measured in terms of abundance, biomass and length. In the present work, only presence and absence data were considered for roach because in a few sites there was insufficient information about abundance as well as biomass of roach.

Model development

Performance criteria to assess of CTs were based on the percentage of correctly classified instances (CCI %) and the Cohen's Kappa statistic (*K*, Cohen, 1960) :

CCI= $(a + d)/N$

Cohen's Kappa statistic=

$$\frac{\left[(a+d) - (((a+c)(a+b) + (b+d)(c+d))/N)\right]}{\left[N - (((a+c)(a+b) + (b+d)(c+d))/N)\right]}$$

where a is true positive (TP), b is false positive (FP), c is false negative (FN) and d is true negative (TN). This is derived from confusion matrices (Fielding and Bell, 1997). By doing this, it is possible to tabulate the presence/absence of roach against the predicted values. CCI is calculated as the percentage of the true positive and true negative predictions. K is a derived statistic that measures the proportion of all possible cases of the presence or absence that are predicted

correctly by a model after accounting for chance predictions. In this study, models with CCI higher than 70% and K higher than 0.40 were considered reliable (D'heygere *et al.* 2006; Dakou *et al.* 2007; Goethals *et al.* 2007; Hoang *et al.* 2010).

Model training and validation were based on a 3-fold cross-validation technique in which, two-third of data were allocated to the training and the remaining one-third to the testing or validation sets. Cross-validation is the statistical method to partition a sample of data into subsets such that the analysis is primarily performed on a single subset, while the other subset(s) are held for subsequent use in confirming and validating the initial analysis. This method is often applied for predicting the error rate of a learning algorithm. Therefore the trade-off between the size of the subsets for training and validation is of crucial importance and needs to be balanced to ensure that the training and validation are done in a 'globally' optimal manner. In the Weka software (Witten and Frank, 2000), a standard way is based on stratified 10-folds cross-validation. Nevertheless, data or time constraints can make 3 or 5-folds cross-validation more convenient (Goethals, 2005).

The applied modelling techniques to model the presence of roach are classification trees (J48 algorithm) (Breiman *et al*. 1984). These techniques that are often referred to as decision trees (Quinlan, 1986) are efficient tools for the solution of classification and regression problems. Decision tree analysis is one of the main techniques used in so-called data mining. The common ways to perform the algorithms are top-down induction of decision trees (Quinlan, 1986). The C4.5 decision tree algorithm (Quinlan, 1993) is a supervised learning approach of machine learning. The J48 algorithm (Weka.classifiers.trees.J48) is a Java reimplementation of C4.5 (Witten and Frank, 2000). Classification trees categorize variables of a hierarchical decision scheme or multidimensional feature space into classes. In classification trees, a feature is checked by each internal node of a tree and a class or category is assigned by each leaf node and the arcs out of a node are labelled with the possible values of the features of this node. An important aspect in classification tree learning is the amount of branches. When there are many branches, the classification trees are difficult to interpret and often these last branches do not contribute significantly to the reliability of the trees. When modelling species presence/absence, the procedure begins with the entire data set, also called the root node and formulates split-defining conditions for each possible value of the explanatory variables to create candidate splits. Next, the algorithm selects the candidate split that minimizes the misclassification rate and uses it to partition the data set into two subgroups. The algorithm continues recursively with each of the new subgroups until no split yields a significant decrease in the misclassification rate or until the subgroup contains a small number of observations. In the present work, standard settings were used for inducing classification trees (J48 algorithm). The only exception was to use the Pruning Confidence Factors (PCFs). In order to reduce the noise in the data and to improve the predictive results with regard to complexity and accuracy of the predictions, several optimization methods can be applied like pruning, bagging and boosting. Pruning is a labour negations problem in classification trees. The simple classification trees perform better than the more complex ones and it makes more sense too (Witten and Frank, 2000). The confidence factor, which is often used for this purpose, 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. In the present work, the binary splits were set as false (in the standard manner). Optimal 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. The class for generating the constructed trees was set as 'pruned'. On the basis of this, PCFs were tested in 4 levels (0.50, 0.25, 0.10 and 0.01) but PCF 0.01 was merely used because

PCFs with the highest confidence factor (e.g. 0.5 and 0.25) result in a greater complexity of trees and then leading to a more difficult interpretation of the constructed trees.

RESULTS

Performance criteria

The predictive results based on the CCI (%) and *K* of a 3 fold cross-validation are presented in Table 2. As demonstrated here, the results were acceptable in terms of both performance criteria. For the percentage of CCI, the models ranged from 78.79% (subset 1) to 82.83% (for each three subsets). According to the overall means of CCI and *K* thresholds, the presence/absence of roach could be predicted reliably by the CTs because CCI for three subsets showed higher value (80.94% ± 1.68) and the value of *K* ranged from 0.58 (subset 1 with PCF 0.01) to 0.66 (subset 3 with PCF 0.50 and 0.25). The overall mean of *K* for three subsets (with different PCF levels) seemed to be also reliable (0.62 ± 0.03). According to Table 2, one can conclude that the models performed logically well. This demonstrates that the predictions were not only based on chance. Based on the average of the CCI%, more than 80% of instances were correctly classified and their respective *K* was also obtained higher than 0.60. The classification tree models were more complex in particular in subset 2 since the number of leaves ranged from 27 (with PCF = 0.50) to 9 (with PCF = 0.01) and also the size of trees was almost big, ranging from 53 to 17. This could lead to difficulties in the interpretation of models. The second complex model was encountered in subset 1. In that case, the number of leaves ranged from 18 (with PCF 0.50) to 9 (with PCF 0.01). Here, the size of tree ranged from 35 to 17. Based on the given problems, it was required to reduce the complexity of trees and derive a better general rule for the presence/absence of roach. On the basis of this, the PCF of 0.01 was chosen as an optimal confidence level of pruning for the prediction. By doing this, it was possible to get more insight into the constructed tress and have a better interpretation of the obtained results.

Table 2. Comparison of the predictive performance of habitat use of roach

Subsets	PCF	CCI (%)	Mean CCI (%) ± SD	CCI (%) overall mean ± SD	Kappa	Mean Kappa ± SD	Kappa overall mean	Number of leaves (model complexity)	Size of the tree
Subset 1	0.50	82.32	81.31 ± 1.69	80.94 ± 1.68	0.65	0.63 ± 0.03	0.62 ± 0.03	18	35
	0.25	82.32			0.65			18	35
	0.10	81.82			0.64			15	29
	0.01	78.79			0.58			9	17
Subset 2	0.50	82.32	80.18 ± 1.72		0.65	0.61 ± 0.03		27	53
	0.25	80.30			0.61			21	41
	0.10	79.29			0.59			14	27
	0.01	78.28			0.57			9	17
Subset 3	0.50	82.83	81.32 ± 1.75		0.66	0.63 ± 0.03		10	19
	0.25	82.83			0.66			10	19
	0.10	79.80			0.60			5	9
	0.01	79.80			0.60			5	9

J48 algorithm

Fig. 2 shows an example of a J48 pruned tree for the prediction of the habitat use of roach in the rivers. The trees revealed interesting information concerning the variables that were important to predict this fish species. The main predictors for the prediction of roach were some physical-habitat variables (e.g. distance from the source and width) since the trees started with the given variables. From the values in the leaves of the trees, one can conclude that many instances were assigned to distance from the source and width. According to the derived rules (based on IF and THEN), if distance from the source was ≤ 6.5 km then roach were absent while no precise decision was made when the distance from the source reached higher than 6.5 km because it still depended on other variables. When width was > 5.5 m, roach population seemed to

be present in the wide river. Most water quality variables didn't play an important role for the prediction of habitat use of roach in rivers so that for these variables only a few instances were assigned to the class presence/absence of roach. For instance, the rules were too weak for the pH. And also when electric conductivity was lower or higher than 857 µS/cm, it had the lowest importance in the constructed trees but the rules were more or less acceptable when the electric conductivity reached lower than 1033 µS/cm.

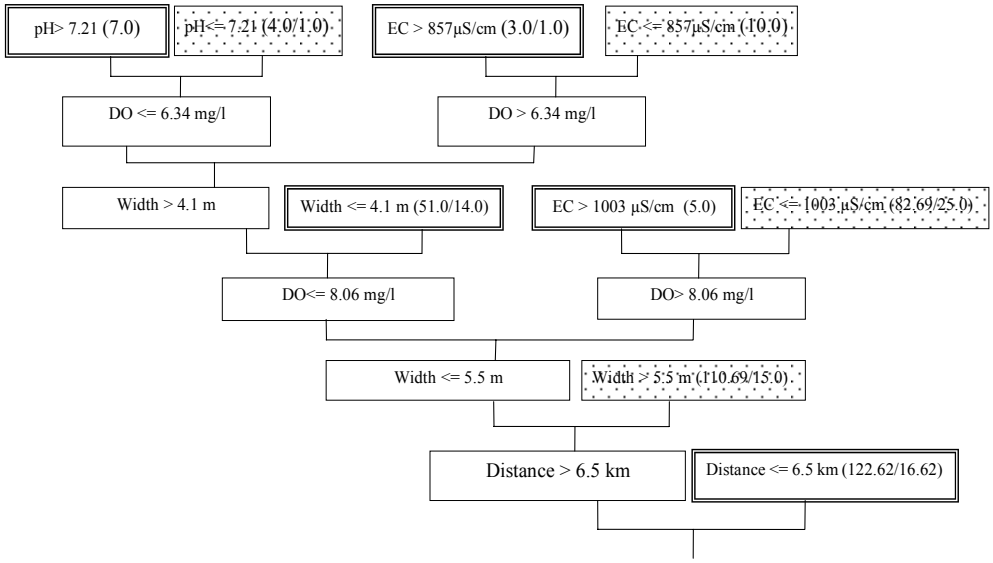

Fig 2. A J48 pruned tree (subset 1) with PCF 0.01 for the prediction of the presence/absence of roach by using input environmental variables (between brackets the number of correct/incorrect predictions in the validation set is mentioned). Double frames rectangles contain the 'absent' of roach while the dotted rectangles denote the 'present' of roach. The third type of rectangle (the simplest one) represents the variables that have not been decided yet for the prediction of roach. W.T°C: water temperature, Distance: distance from the source, DO: dissolved oxygen and EC: electric conductivity.

DISCUSSION

A combination of two model performances expressed as percent of correctly classified instance (CCI%) and *K* seemed to have better outcomes to predict the most important variables for the habitat use of roach in the rivers because measuring the predictive ecosystem model performance frequently entails calculating the percentage of the sites for which presence/absence of organisms (e.g. roach in this study) is correctly predicted (Manel *et al.,* 2001). A logical relationship was observed between CCI and *K* in the models when applying different pruning confidence factors. This can most likely be explained that frequency of roach occurrence was equal in all sampling sites (50% as absence and 50% as presence). Several authors (e.g. Fielding and Bell, 1997; Manel *et al.,* 1999; Goethals and De Pauw, 2001; Dedecker *et al.,* 2002; D'heygere *et al.,* 2003; Dakou *et al.,* 2006) explored that frequency of occurrence of organisms influence model performances. Based on the obtained results, the prediction outcomes were satisfactory for roach so that majority of PCFs showed a reliable prediction between predicted and observed outputs. In other words, there wasn't a big difference based on the two performance criteria. As a result, classification trees (J48) could predict well the occurrence of roach in rivers. Olden and Jackson 2002 found the similar results when applying classification trees and neural networks for predicting the presence/absence of 27 fish species including roach. According to authors, non-linear modelling methods (e.g. classification trees and neural networks) have the capability to capture and model complex, non-linear patterns which are observed in ecological data. D'heygere *et al.* (2003) and Dakou *et al.* (2006) developed similar

modelling methods in Flanders to predict macroinvertebrates taxa and reached to a similar conclusion, demonstrating that classification trees systems select the most promising attribute to split at each branch of a tree. On the other hand, Dzeroski *et al.* (1997) explored that the decision and regression trees are applied to ascertain the ecological needs of organisms, which might be difficult to realize. In the present study, the main weakness of the induced trees, however, was the complexity of the constructed trees and the variables appearing on them. In most subsets, the trees consisted of many variables especially when the highest PCFs (e.g. 0.50) were applied. In such a way, it was difficult to find a general rule in the models. That was the main reason that PCFs were applied in the lowest value to realize the most explanatory predictors for the prediction of habitat use of roach.

In the present research, only presence and absence data were considered for the prediction of habitat use of roach. The reason was that in some sites, there was insufficient information about abundance and biomass of roach. Some important predictors (e.g. depth and slope) were not directly displayed in the lower PCFs (e.g. 0.01 and 0.10) but appeared in the higher PCFs (those PCFs were not mentioned in the results). For instance, the trees showed that in non-deep waters roach were present (lower than 0.70 m). Brosse and Lek (2000) showed that the most important variables influencing the 0+roach distribution was distance from the bank, depth, local slope of the bottom, percentage of mud and flooded vegetation cover. The distance from source played the key role for the prediction of presence/absence of roach followed by width. The variable of distance from the source reveals that roach populations prefer to inhabit the downstream part of rivers. Brabrand and Faafeng (1994) and Garner (1995) stated that depth constitutes an essential feature in 0+roach habitat preference considering the needs for shelters against predation. In another study conducted by Brabrand and Faafeng (1994) and Eklov (1997), it was also revealed that roach avoids deep waters and steeply sloping parts because these areas are usually occupied by some top predators e.g. perch (*Perca fluviatilis* L.) and pike (*Esox lucius*). Poizat and Pont (1996) moreover stated that after having removed the variation elucidated at the two larger habitat scales, only depth and shelter showed a significant effect on fish abundance at the microhabitat scale. Copp (1992) explored that habitat use by roach varies in lakes and streams, where current velocity powerfully influences roach habitat. In another study (Garner, 1995; Rossier *et al.,* 1996), it was stressed that roach is strongly associated with aquatic vegetation. Missing values in the dataset and unmeasured input variables resulted in to eliminate some valuable inputs variables e.g. vegetation covers, flow velocity, habitat quality and etc. Therefore more reliable prediction for habitat use of roach would be obtained if there were adequate and more relevant structural-habitat variables. In addition to this, another drawback of the current study was the number of samples taken in the fields. So the samples were not equally distributed over different river basins and Huet zones. Unequal sampling could certainly have an undesirable effect on the outcomes of models as well as prediction because some sites were sampled repeatedly and others were not.

Water temperature was never recognized as an important predictor by the induced trees. With looking at the mean temperature, one can realize that water temperature does not fluctuate a lot in the river basins situated in Flanders. Otherwise, it was necessary to realize how seasonality would affect the habitat use of roach in the rivers. However, pH and dissolved oxygen appeared on the trees but contributed less to the prediction of roach. A possible reason could be that roach is a dominant fish species under eutrophic conditions (Persson, 1983). They have the capability to capture many small zooplanktons and can utilize primary producers as food resources (Johansson and Persson, 1986; Persson and Greenberg, 1990) and also have the ability to feed on dead organic matter (Persson, 1983; Persson and Greenberg, 1990; Vinni *et al.,* 2000). The only important water quality variable was electric conductivity; nevertheless it was detected in the end of trees, demonstrating to contribute less to

the prediction of roach in rivers. Some important water quality variables were missing in the dataset such as $N\text{-}NO_3^-$, $N\text{-}NH_4^+$, $P\text{-}PO_4^{3-}$ and etc. while these variables represent nutrient contents in most water courses. If they were introduced to the model, they might have contributed more to the prediction of roach.

ACKNOWLEDGEMENTS

The authors wish to thank the Institute for Forestry and Game Management for data collection.

REFERENCES

Belpaire, C., Smolders, R., Vanden Auweele, I., Ercken, D., Breine, J., Van Thuyne, G. and Ollevier, F. (2000) An Index of Biotic Integrity characterizing fish populations and the ecological quality of Flandrian water bodies. *Hydrobiologia*. **434**, 17-33.

Brabrand, A. (1985) Food of roach, *Rutilus rutilus* and ide, *Leuciscus idus*: significance of diet shift for interspecific competition in omnivorous fishes. *Oecologia*. **66**, 461-467.

Brabrand, A. and Faafeng, B. (1994) Habitat shift in roach, *Rutilus rutilus* induced by the introduction of pike-perch, *Stizostedion lucioperca*. *Limnologie*. **25**, 21-23.

Breine, J., Simoens, I., Goethals, P.L.M., Quataert, P., Ercken, D., Chris, V. L. and Belpaire,C. (2004) A fish-based index of biotic integrity for upstream brooks in Flanders (Belgium). *Hydrobiologia*. **522**, 133-148.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) Classification and Regression Trees. Wadsworth, Pacific Grove, CA, USA.

Brosse, S. and Lek, S. (2000) Modelling roach, *Rutilus rutilus* microhabitat using linear and nonlinear techniques. *Freshwater. Bio.* **44**, 34-41.

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

Copp, G. H. (1990) Shifts in the microhabitat of larval and juvenile the roach, *Rutilus rutilus* L. in a floodplain channel. *J. Fish Biol*. **36**, 683-692.

Copp, G. H. (1992) An empirical model for predicting microhabitat of 0+ juvenile fishes in a lowland river catchment. *Oecologia*. **91**, 338-345.

Dakou, E., Goethals, P.L.M., D'heygere, T., Dedecker, A.P., Gabriels, W. and De Pauw, N. (2006) Development of artificial neural network models predicting macroinvertebrate taxa in the river Axios (Northern Greece). *Annales de Limnologie-Ann. Limnol.- Int. J. Lim*. **42**, 241- 250.

Dakou, E., D'heygere, T., Dedecker, A.P., Goethals, P.L.M., Lazaridou-Dimitriadou, M. and De Pauw, N., (2007) Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquat. Ecol*. **41**,399-411.

Dedecker, A.P., Goethas P.L.M., Gabriels, W. and De Pauw, N. (2002) Comparison of Artificial Neural Network (ANN) model developments methods for prediction of macroinvertebrates communities in the Zwalm river basin in Flanders, Belgium. *The ScientificWorldJo.* **2**, 96-104.

D'heygere, T., Goethals, P. L. M. and De Pauw, N. (2003) Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinverteberates. *Ecol. Model*. **160**, 291-300.

D'heygere, T., Goethals, P. L. M. and De Pauw, N. (2006) Genetic algorithms for optimization of predictive ecosystems models based on decision trees and neural networks. *Ecol. Model*. **195**, 20-29.

Dzeroski, S., Grobovic, J. and Walley, W.J. (1997) Machine learning applications in biological classification of river water quality, pp.429-448. In: Michalski, R.S., Bratko, I. & Kubat, M. Machine learning data mining: methods and applications. John Wiley and Sons Ltd., New York.

Eklov, P. (1997) Effects of habitat complexity and prey abundance on the spatial and temporal distributions of perch, *Perca fluviatilis* and pike, *Esox lucius*. *Can. J. Fish. Aquat. Sci.* **54**, 1520-1531.

Fielding, A.H. and Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv*. **24**, 38-49.

Garner, P. (1995) Suitability indices for

juvenile 0+ roach, *Rutilus rutilus* (L.) using point abundance sampling data. *Regul. River*. **10**, 99-104.

Goethals, P.L.M. and De Pauw, N. (2001) Development of a concept for integrated ecological river assessment in Flanders, Belgium. *J. Limnol*. **60**, 7-16.

Goethals, P. L. M. (2005) Data driven development of predictive ecological models for benthic macroinvertebrates in rivers. PhD thesis. University of Ghent. 377 pp.

Goethals, P.L.M., Dedecker, A.P., Gabriels, W., Lek, S. and De Pauw, N. (2007) Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquat. Ecol*. 41, 491-508.

Horppila, J. (1994) The diet and growth of roach, *Rutilus rutilus* (L.)) in Lake Vesijarvi and possible changes in the course of biomanipulation. *Hydrobiologia*. **294**, 35-41.

Kahl, U., Dorner, H., Radke, R.J., Wagner, A. and Benndorf, J. (2001) The roach population in the hypertrophic Bautzen Reservoir: structure, diet and impact on Daphnia galeata. *Limnologica*. **31**: 61-68.

Kahl, U. and Radke, R. J. (2006) Habitat and food resource use of perch and roach in a deep mesotrophic reservoir: enough space to avoid competition? *Ecol. Freshw. Fish*. **15**, 48-56.

Jackson, D.A. and Harvey, H.H. (1997) Qualitative and quantitative sampling of lake fish communities. *Can. J. Fish. Aquat. Sci*. **54**, 2807-2813.

Johansson, L. and Persson, L. (1986) The fish community of temperate, eutrophic lakes. In: Riemann, M.B.S. ed. Carbon dynamics of eutrophic, temperate lakes: the structure and functions of the pelagic environment. Amsterdam: Elsevier, pp. 237-266.

Lawton, J. (1996) Patterns in ecology. *Oikos*. **75**, 145-147.

Hoang, T.H., Lock, K., Mouton, A. and Goethals, P. L.M. (2010) Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecol. Inform.* **5**, 140-146.

Manel, S., Dias, J.M., Buckton, S.T. and Ormerod, S. J. (1999) Alternatives methods for predicting species distribution: an illustration with Hialayan river birds. *J. Appl. Ecol.* **36**,734-747.

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

Olden, J.D. and Jackson, D.A. (2002) A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biol.* **47**, 1976-1995.

Quinlan, J.R. (1986) Induction of decision trees. *Mach.Learn.* **1(1)**, 81-106.

Quinlan, J.R. (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco, USA.

Persson, L. and Greenberg, L.A. (1990) Juvenile competitive bottlenecks- the perch, *Perca fluviatilis*- roach, *Rutilus rutilus* interaction. *Ecology*. **71**, 44-56.

Persson, L. (1983) Effects of intraspecific and interspecific competition on dynamics and size structure of a perch, *Perca fluviatilis* and a roach, *Rutilus rutilus* population. *Oikos*. **41**,126-132.

Poizat G. and Pont D. (1996) Multi-scale approach to species-habitat relationships: juvenile fish in a large river section. *Freshwater Biol.* **36**, 611-622.

Ricciardi, A. and Rasmussen, J.B. (1999) Extinction rates of North American freshwater fauna. *Conserv. Biol*. **13**, 1220-1222.

Rossier, O., Castella, E. and Lachavanne, J.B. (1996) Influence of submerged aquatic vegetation on size class distribution of perch, *Perca fluviatilis* and roach, *Rutilus rutilus* in the littoral zone of Lake Geneva (Switzerland). *Aquat. Sci*. **58**, 1-14.

Schoener, T. (1974) Resource partitioning in ecological communities. *Science*. **185**, 27-39.

Schulze, T., Dörner, H., Hölker, F. and Mehner, T. (2006) Determinants of habitat use in large roach. *J. Fish Biol.* **69**, 1136-1150.

Sharma, C.M. and Borgstrøm, R. (2008) Shift in density, habitat use, and diet of perch and roach: An effect changed predation pressure after manipulation of pike. *Fish. Res.* **91**, 98-106.

Skov, C., Berg, S., Jacobsen, L. and Jepsen, N. (2002) Habitat use and foraging

success of 0+ Pike, *Esox lucius* (L.) in experimental ponds related to prey fish, water transparency and light intensity. *Ecol. Freshw. Fish*. **11**, 65-73.
Vinni, M., Horppila, J., Olin, M., Ruuhijarvi, J. and Nyberg, K., (2000) The food, growth and abundance of five co-existing cyprinids in lake basins of different morphometry and water quality. *Aquat. Ecol*. **34**, 421-431.
Werner, E.E., Hall, D.J., Laughlin, D.R., Wagner, D.J., Wilsmann, L.A. and Funk, F.C. (1977) Habitat partitioning in a freshwater fish community. *J. Fish.Res.Board.Can*. **34**, 360-370.
Witten, J.H. and Frank, E. (2000) Data mining: practical machine learning tools and techniques with Java implementations, Morgan Kaufman publishers, San Francisco. 369 pp.

(Received: Jul. 5-2010, Accepted: Nov. 22-2010)

کاربرد روش کلاسه بندی درختی(خوشه ای) برای مدل کردن حضور (یا عدم حضور) ماهی کلمه در رود خانه ها

ر. زرکامی

چکیده

در مقاله حاضر با استفاده از مدل کلاسه بندی خوشه ای میزان پراکنش ماهی کلمه Rutilus rutilus در رودخانه‌های منطقه فلاندر در بلژیک مطالعه شده است. با توجه به میزان فراوانی ماهی و خصوصیات ساختاری و فیزیکی-شیمیایی محیط مورد مطالعه حضور و عدم حضور ماهی مدل شده است. برای بررسی اعتبار و کیفیت مدل از ۲ شاخص آماری مهم استفاده شده است:۱- شاخص کاپای کوهنیk ۲- تعداد داده های صحیح کلاسه بندی شده (CCI). با استفاده از روش هرس pruning confidence factor که در ۴ سطح مختلف آماری ,0.5 (0.25, 0.1, 0.01) انجام شده و همچنین با به کار گیری ۳ بار تکنیک‌های cross-validation بهترین k و CCI مشخص شده است. نتایج حاصله نشان داد که مدل از ضریب اعتماد معقولی برخوردار بوده و این امر به در صد بالای پیش بینی ماهی کلمه منجر شده است. بر اساس نتایج این مدل، هم ویژگی‌های ساختاری محیط و هم خصوصیات کیفی آب در پراکنش ماهی کلمه نقش دارند اما فاکتورهای های ساختاری محیط بیش از فیزیکی و شیمیایی در پراکنش کلمه تاثیر می‌گذارند. بر اساس مدل توسعه داده شده ۲ تا از فاکتورهای مهم در این راستا فاصله رودخانه از منابع بالا دست و عریض بودن رودخانه می‌باشند. نظر به این که ماهی کلمه از جمله ماهیانی است که تحملش به شرایط نامسأعد محیط زیاد است وجود چنین ارتباطی بین ساختار محیط با حضور و عدم حضور ماهی کاملا منطقی است.