Caspian J. Env. Sci. 2010, Vol. 8 No.1 pp. 55~63

CJES
Caspian Journal of Environmental Sciences

[Research]

Use of classification tree methods to study the habitat requirements of tench (*Tinca tinca*) (L., 1758)

R. Zarkami[1]*, P. Goethals[2] and N. De Pauw[2]

1- Department of Environmental Sciences, Faculty of Natural Resources, University of Guilan, P.O. Box 1144, Sowmeh Sara, Guilan, Iran.
2- Department of Applied Ecology, Ghent University, J. Plateaustraat 22, B-9000 Gent, Belgium.
** Corresponding author's E-mail: rzarkami2002@yahoo.co.uk*

ABSTRACT
Classification trees (J48) were induced to predict the habitat requirements of tench (*Tinca tinca*). 306 datasets were used for the given fish during 8 years in the river basins in Flanders (Belgium). The input variables consisted of the structural-habitat (width, depth, gradient slope and distance from the source) and physic chemical (pH, dissolved oxygen, water temperature and electric conductivity), and the output ones were the abundance and presence/absence of tench. To find the best performance model, a three-fold cross validation was applied on the entire dataset. In order to evaluate the model stability, the dataset were remixed in 5 times, obtaining in total 15 different model training and validation events. The effect of pruning on the reliability and model complexity was tested in each subset. The performance evaluation was based on a combination of the number of Correctly Classified Instances (CCI) and Kappa statistic. The results showed that the predictive performance evaluation was suitable, confirming the reliability of classification trees methods. The overall average of CCI and Kappa for the prediction of tench was obtained 75.8% and 0.53. When analyzing the ecological relevance of classification trees, it seemed that the structural-habitat variables were important predictors compared to physic chemical variables.

Keywords: *tench (Tinca tinca), classification tree models (J48), physical-chemical variables, structural-habitat variables, Flanders river basins*

INTRODUCTION
During the last decades, many fish species in Flanders (Belgium) decreased enormously; therefore the structural quality of the habitat is often too poor to support a diverse and balanced fish population, in particular in the canals (Belpaire, 2000). A key issue in conservation management and river restoration is to get acquainted with the relationship between the environmental factors and the occurrence of the freshwater organism. In this perspective, modeling techniques are becoming as important tool to support decision-making in water management and conservation. Prediction of organisms by modeling techniques has been an interesting subject for many researches (Gaston and Blackburn, 1999; Olden and Jackson, 2002; Goethals *et al*., 2002; Dedecker *et al*., 2002; D'heygere *et al*., 2003; Dakou *et al*., 2006 *a, b*). Among these, decision tree (Quinlan, 1993) is well known as a powerful tool to predict freshwater organisms. Predictive models play significant role in environment conservation, biological monitoring and resource assessment (Fielding and Bell, 1997). Moreover, the given models are known as the alteration and loss of aquatic habitat and also as the core factor threatening the conservation of fish populations and communities (Richter *et al*., 1997; Harig and Bain, 1998; Ricciardi and Rasmussen, 1999). When predicting the presence/absence of organisms by models, one has to pay more attention to the collection of suitable model inputs (Kaastra and Boyd, 1995; Faraway and Chatfield, 1998).

Tench are Cyprinid fish, belonging to the same family of goldfish, rudd and koi carp. They are wide spread in European countries and neighboring regions. Tench populations are not important as commercial fish because they have slow growth rate and tasteless flesh (Yilmaz, 2002). This study mainly aimed to develop classification tree models (J48) to predict the habitat requirements of tench in the river basins in Flanders.

MATERIAL AND METHODS

Study area

Flanders is located in the northern part of Belgium and has several major river basins (Fig. 1). The main river basin covering nearly whole Flanders is the Scheldt (which is about 70% of Flanders). For management aims the Scheldt River basin itself is divided in 8 subbasins: Upper Scheldet, Leie, Dender, Nete, Demer, Dijle, Polder and canals around Ghent.

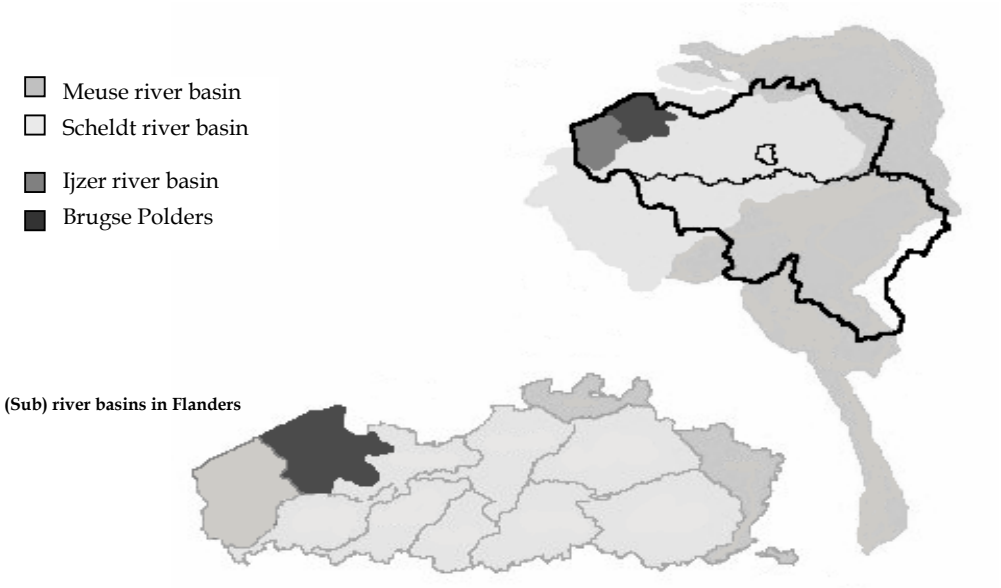

Fig. 1. Location of Flanders in Belgium (the Scheldt and Meuse river basins in France, Belgium and the Netherlands. Major part of the Scheldt river basin is located in Flanders (Goethals, 2005).

Database set-up

The dataset consisted of measurements of 306 instances for tench collected in 8 river basins in Flanders. At each sampling sites, several environmental variables were recorded during monitoring campaigns (Table 1).

Table 1. Biotic and abiotic input variables used for the prediction of the habitat suitability of tench in the river basins in Flanders.

River characteristics	Unit	Minimum	Maximum	Mean ± SD
Water temperature	˚C	2.5	23.1	10.74 ± 3.90
Distance from the source	Km	0.0	84.8	20.50 ± 20.90
Width	m	0.4	66.7	7.00 ± 6.60
Slope	%	0.0	35	2.30 ± 3.50
Depth	m	0.1	2.5	0.65 ± 0.45
Dissolved oxygen (DO)	mg l^{-1}	1.3	14.9	8.30 ± 2.30
pH		5.3	8.6	7.30 ± 0.50
Electric conductivity (EC)	□S/cm	153	5220	787 ± 575
Abundance	N	0.0	1200	61 ± 152

SD: standard deviation

These measurements consisted of the physic chemical variables: electric conductivity, dissolved oxygen, pH, water temperature and the structural-habitat variables: gradient slope, distance from the source, width, and depth. Biotic

variable was only the abundance of tench in which the tench presence-absence was obtained from this variable. Abundance data was also used for visualizing the scatter plot of physico-habitat variables for tench. The abiotic variables were used as input variables and biotic variables as output in the J48 model included in the Weka toolbox (Witten and Frank, 2000). Before model development it was important to determine the frequency of occurrence of tench in the sites. The frequency of occurrence (observed values) for tench was considered 50% in all monitored sites of which tench were present in 153 cases (50%) and were absent in 153 cases (50%). The geographical distribution of presence/absence of the given species in the monitored sites is visualized in Fig. 2.

Electric-fishing method was conducted with a 5 KW generator (with voltage of 300/500V and a pulse frequency of 480 Hz). The number of hand-held anodes was 2. Further information on sampling methodologies is given by Belpaire *et al*. (2000). In this monitoring approach, to use the classification tree methods, if no fish species were caught the number of "0" (as absence) otherwise number of "1" (as presence) was represented.

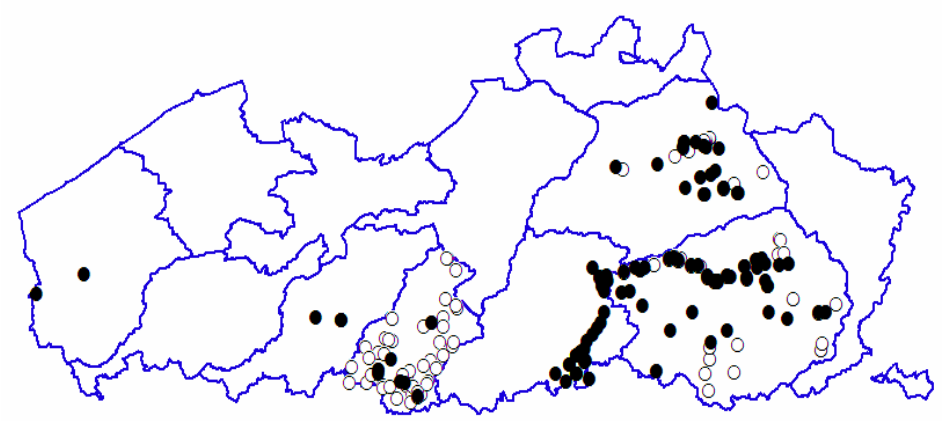

Fig 2. Geographical distribution of tench in the river basins in Flanders (the fish presence is indicated as filled circles and absence is indicated as open circles).

Classification trees

Classification trees (Breiman *et al*., 1984), often referred to as decision trees (Quinlan, 1986) that predict the value of a discrete dependent variable with a finite set of values (called class) on the basis of the values of a set of independent variables (called attributes), which may be either continuous or discrete. In this paper, the dependent variables were the biotic variable (the presence/absence of tench) and the independent variables were the 8 abiotic variables listed in Table 1. The common way to induce classification trees is Top-Down Induction of classification trees that starts with the entire set of training examples (Quinlan, 1986). The Classification trees system recursively partitions the dataset into smaller subsets by selecting one attribute to 'branch' on (creating one or more most informative attribute is selected by introducing a function that assigns a value of the quality of the partition obtained by a specific attribute. For discrete attributes, a branch of the tree is typically created for each possible value of the attribute. For continuous attributes, a threshold is selected and two branches are created based on that threshold. The final subsets formed by the recursive process are called the 'leaves' of the decision tree and each tree is labelled with a class. In this study, classification tree (J48) was used, which is a Java re-implementation of C4.5 (Quinlan, 1993) and is a part of the machine learning package Weka. Here, J48 with default values of the parameters was used for inducing classification tree. The only exception was the use of the Pruning Confidence Factor (PCF) in which PCF 0.1 was applied for tench.

Model training and validation

The prediction accuracy of the induced model was evaluated based on two model indicators: the percentage of CCI (Correctly Classified Instances) and the Cohen's Kappa statistic (Cohen, 1960). Model training and validation for the fish species was applied according to a 3-fold cross-validation. Here, 204 instances (two-third of datasets) were allocated for the training and 102 instances (one-third) for the validation set respectively. The stability of the model development was tested on the basis of five independent remixes as well as three subsets in the dataset. Due to simplicity in interpretation of results only 1 out of 5 remixes with high reliability was considered for the fish species.

Table 2. Predictive results of classification tree models based on J48 with pruning optimization (the third time data randomization was considered for tench)

Fish species	Frequency of occurrence (%)	SSi	CCI (%)	Mean CCI (%) ± SD	Kappa	Mean Kappa ± SD	Number of leaves (model complexity)
Tench		SS1	80.4		0.60		8
(remix 3)	50	SS2	74.5	75.8 ± 4.1	0.50	0.53 ± 0.08	15
		SS3	72.6		0.50		7

SSi: the number of subsets and SD: standard deviation

Among 3 subsets the highest predictive results were obtained in the subset 1 (CCI 80.4% and Kappa 0.60). In Fig. 3, the predictive results based on CCI and Kappa is visualized for tench. As a result, the reliability of performance model for the prediction of tench seems to be good. The number of leaves ranged between 7 (subset 3) and 15 (subset 2) which increased the complexity of trees, hence caused the ecological interpretation very difficult.

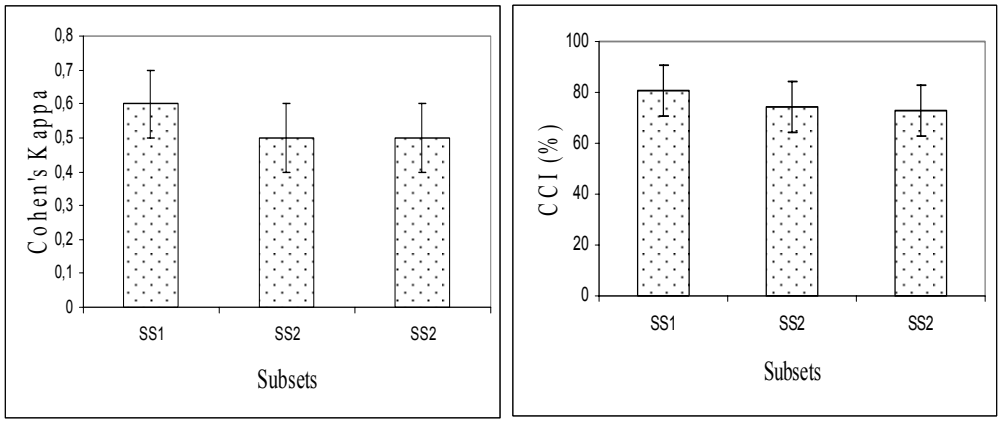

Fig. 3. Evaluation of classification trees (J48) based on CCI% (a) and Cohen's Kappa (b) for tench in the river basins in Flanders (SSi: Subsets).

The induced classification tree models

The constructed trees (J48) for tench are presented in Fig. 4. From the given trees, general rules (in terms of IF and THEN) can be deduced for the presence/absence of tench. These rules can be derived from the leaf to the root of the trees. As visualized here, the induced trees grew with 6 leaves (size of the trees 11). For simplicity in the interpretation of the results, only the remix 3 from subset 2 was selected. In order to reduce the complexity of the trees, PCF 0.1 was merely applied. In the given subset and remix, more reliable prediction was obtained so that 79 of 102 instances (77.5%) were correctly classified with the respective Kappa 0.55 (the number of instances is not directly indicated for each variable displayed in figure 4). As seen in the given trees, all 4 structural-habitat variables and also one physico-chemical variable (water temperature) were detected for the prediction of presence/absence of tench. In subset 1 and 3, the structural-habitat variables (particularly distance from the source) were also the most dominant variables for the prediction of the presence-absence of tench. Pearson correlation coefficient (p<0.05) calculated between the structural-habitat variables and abundance of tench revealed that the individuals were positively correlated with distance from the source (r = 0.30), width (r = 0.25), depth (r = 0.23) and negatively correlated with slope (r = -0.22). In relation to the physic chemical variables, tench populations were weakly and negatively correlated with dissolved oxygen (r = -0.14) and electric conductivity (r = -0.01) and positively with pH (r = 0.27) and water temperature (r = 0.15). When distance from the source is approximately lower than 16 km and the slope is lower than 0.6%, tench populations were present. On the contrary, in slope higher than 0.6%, they were missing. When distance from the source reached higher than 16 km and river width became lower than 8 m, tench populations were present while in the river width of higher than 8 m and deep river of higher than 2 m they were absent. In shallow water (lower than 2 m) with water temperature of lower than 7°C they were not found. Fig. 5 illustrates the scatter plot representing the relationship between distance from the source and abundance of tench (this figure represents a particular example of structural-habitat variable). As visualized here, the increase of the distance from the source leads to the decease of abundance of tench populations.

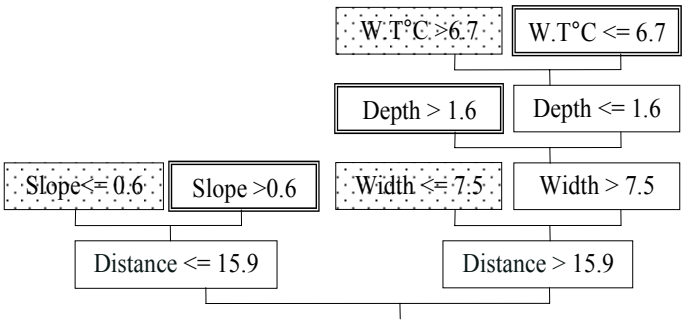

Fig. 4. J48 pruned trees of tench in the sampling sites in Flanders (an example of subset 2 from remix 3 with PCF 0.1, double frames contain absent of tench while the dotted rectangles contain present of tench, W.T°C: water temperature, Distance: distance from the source).

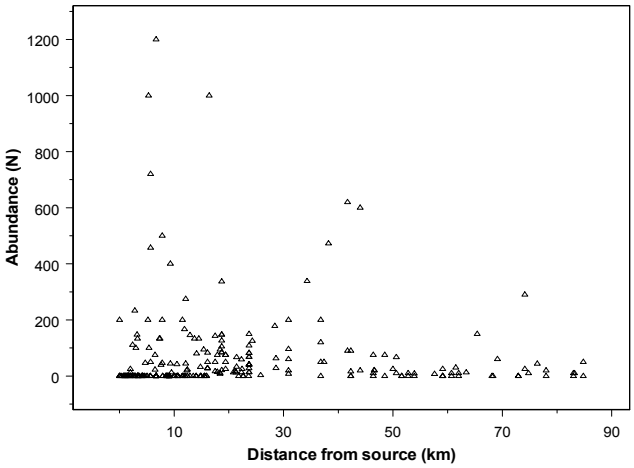

Fig. 5. Scatter plot representing the relationship between distance from the source and abundance of tench in the river basins in Flanders (only 285 instances are visualized).

DISCUSSION

In the present study, the use of classification tree models to get insight in the habitat preferences of tench seemed to be successful in terms of CCI and Kappa statistic. As frequency of occurrence of tench was equal in the sampling sites (50%), a logical relationship was obtained for the prediction of tench based on these two model performances. Several authors reached the same conclusions (e.g. Manel *et al*., 2001; Goethals *et al*., 2002; Dakou *et al*., 2006 *a, b*). The authors stated that the predictive performance of classification tree models is strongly related to the frequency of occurrence of the organism. In this study, the model performance for the prediction of tench was acceptable, probably because more and better data were available for the given species. This was in line with in the study of Dakou et al. (2006*a*) demonstrating the close relationship between the frequency of occurrence of the predicted organisms in the dataset and the predictive performance of the models.

When analyzing the ecological relevance of classification trees for tench populations, one can see that the structural-habitat variables were the most dominant ones for the prediction of the habitat suitability of tench in the river basins. Unlike the structural-habitat variables, classification trees didn't present reliable predications for the habitat preferences of tench in terms of all physico-chemical variables. The only explanatory variable for the prediction of presence/absence of tench was water temperature. One of the significant predictors affecting the quantity of habitat preferences of the given fish is water temperature (Casselman *et al*., 1996). Nevertheless this variable was detected in the end of trees. Water temperature had a less contribution to the prediction of tench (compared to the structural-habitat variables) nevertheless the constructed trees were reliable for water temperature. Tench are well-known as warm water fish species and unlike salmonidae (cold water species) prefer temperature over 20°C. Their favorite temperature is 20-21°C and their final preference would be 27.4 ± 0.5°C (Pereze Regadera *et al*., 1994). A preferred range of water temperature for tench is 15-23.5 °C and growth can happen over the range 12-30°C. Tench spawning is also closely dependent on water temperature, but the temperature in relation to spawning differs (Rowe, 2004). Gray and Daule (2001) pointed out that spawning happened in late spring when the range of water temperature was between 10 and 20°C. Low variations of water temperature in Flanders's climate are probably the main reason for appearing this variable in the end of induced trees, confirming the less contribution of this variable to the prediction of habitat requirements of tench in Flanders.

Based on the trees, tench populations avoid inhabiting higher depth (deeper than 1.6 m). This is in line with the study of Rowe (2004), demonstrating that tench are mainly found in shallow, still and slow-moving freshwater environment. However, spawning of tench happens in shallow waters (usually <1 m deep) and tench are able to spawn in a broad range of water temperature, laying their eggs over aquatic vegetation such as macrophyte and reeds (Rowe, 2004). Habitat complexity and food accessibility and are two factors to describe the high density of fish in vegetated habitats (Grenouillet and Seip, 2002), in particular tench populations are very closely dependent on vegetation cover.

In addition to depth, gradient-slope, distance from the source and wetted-width were determined as important predictors for the habitat requirements of tench in the river basins. Classification trees model clearly showed that when the river slope increases, tench tend to be absent. Rowe (2004) reported that a maximum of water velocity for tench is 0.27 m/s. This demonstrates that tench avoid high-gradient slope and rapid water. It is well documented that adult tench inhabit a range of waters typically dominated by low water velocities, soft substrates (e.g. mud, silt or sand) and existence of some aquatic vegetation. Some of such habitats in rivers for tench include the lower reaches of rivers, off-river habitats such as oxbows and river deltas (Donnely et al., 1984; Gonzalez et al., 2000). Based on the classification trees, the distance from the source is the first and main explanatory variable describing habitat requirements of tench; that is why the trees first emerged with this variable. Distance from the source has also a close relation with the gradient-slope as visualized already in the emerged trees. From ecological perspectives, these variables seem to be quite logical and the relevance of these (integrating) structural-habitat variables is confirmed in the literature (Kerle et al., 2001). The variable 'distance from source' illustrated that most tench individuals are found in upstream part of rivers. In essence, the upstream part of rivers is not a suitable habitat for tench individuals. This may be related to water pollution in the downstream part of rivers. If the downstream of rivers was not polluted, many tench individuals would inhabit this part. This was in accordance with the study of Brosse et al. (1999), demonstrating that fish were mainly found in downstream part of river. Tench were found to be present when the river width was lower than 7.5m, while in width higher than 7.5m the absence of tench was mainly related to river depth.

Dissolved oxygen, pH and electric conductivity were never classified as important variables for the given species. This may be explained by the fact that tench are highly tolerant of low oxygen levels (Vainikka, 2003) and can survive in waters whose oxygen levels are as low as 0.7 mg/l (Rowe, 2004), therefore they can somewhat bear the eutrophicated conditions. As a result, the given variable was not taken into account as an important variable for the prediction of presence/absence of tench. Adult tench can highly tolerate a broad range of pH but they prefer the range of pH 6.5-8 (Rowe, 2004). Mortality increases at levels below 5 and over 10.8. The range of observed values of pH in the river basins in Flanders never exceeded 10.8 or dropped 5, demonstrating the less importance of pH for the prediction of tench in the sampling sites.

Although the models were more or less reliable, the prediction of tench could be further improved by including more variables in the standard monitoring network. In particular, the including of some relevant habitat variables (e.g. pool-riffle patterns and dense vegetation cover) can be useful depending on rich vegetation for food collection and regeneration.

ACKNOWLEDGMENTS

The author wishes to thank the Flemish Environment Agency for biotic data and Institute for Forestry and Game Management for chemical (Brussels) data collection.

REFERENCES

Belpaire, C., Smolders, R., Vanden A.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.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and regression trees*, Pacific Grove, Wadsworth.

Brosse, S., Guegan, J.F., Tourenq, J.N. and Lek, S. (1999) The use of artificial neural network to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake, *Ecol. Mod*. **120**, 299-311.

Casselman, J.M. and Lewis, C.A. (1996) Habitat requirements of northern pike (*Esox lucius L.*), *Canadian J. Fish. Aqu. Sci.*. **53,** 161-174.

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

Dakou, E., Goethals, P.L.M., D'heygere, T., Dedecker, A.P., Gabriels, W. and De Pauw, N. (2006*a*) Development of artificial neural network models predicting macroinvertebrate taxa in the river Axios (Northern Greece), *Animal Limnology*. **5**, 10-17.

Dakou, E., D'heygere, T., Dedecker, A.P., Goethals, P.L.M., Gabriels, W. Lazaridou, M. and De Pauw, N. (2006*b*) Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece), *Aquatic Ecology*. **41**, 399-411.

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

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

Donnely, R.E., Caffrey, J.M. and Tierney, D.M. (1998) Movement of bream (*Abramis brama* (L.)), rudd X bream hybrid, tench (*Tinca tinca* (L.)) and pike (*Esox lucius*) in an Irish canal habitat, *Hydro.* **371/372**, 305-308.

Faraway, J. and Chatfield, C. (1998) Time series forecasting with neural network: a comparative study using airline data, *Appl. Stat.*. **47 (2)**, 231-250.

Fielding, A.H. and Bell, J.F. (1997). A review method for the assessment of prediction errors in conservation presence and absence model, *Env. Cons.*. **24**, 38-49.

Gaston, K.J. and Blackburn, T.M. (1999) A critique for macroecology, *Oikos*. **84**, 353-368.

Goethals, P.L.M., Dedcker, A., Gabriels, W. and De Pauw, N. (2002) Development and application of predictive river ecosystem models based on classification trees and artificial neural networks. *Ecological informatics, Understanding ecology by biologically inspired computation*. (ed. Recknagel), Springer, Berlin: 432 pp.

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

Gonzalez, G., Maze, R.A., Dominguez, J. and Pena, J. C. (2000) Trophic ecology of Tinca Tinca in two different habitats in North-West of Spain, *Cybium.* **24**, 123-138.

Gray, R.H. and Daule, D.D. (2001) Some life history characteristics of cyprinids in the Handford reach, mid-Columbia River, *North. Sci.* **75**, 122-136.

Grenouillet, G.l., Pont, D. and Seip, K.L. (2002) Abundance and species richness as a function of food resources and vegetation structure: juvenile fish assemblages in rivers, *Ecography.* **25**, 641-650.

Harig, A.L. and Bain, M.B. (1998) Defining and restoring biological integrity in wilderness Lakes, *Ecol.l Appl.s.* **8**, 71-87.

Kaastra, I. and Boyd, M. S. (1995) Forecasting future trading volume using neural networks, *Journal of Future Markets.* **15(8)**, 953-970.

Kerle, F., Zollner, F., Kappus, B., Marx, W. and Giesecke, J. (2001) Fish habitat and vegetation modelling in floodplains with CASIMIR. CFR project report 13, IWS, University of Stuttgart, 75 pp.

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

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

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

Quinlan, J.R. (1993) *C4.5: program for machine learning*. Morgan Kaufmann publishers, San Francisco.302 pp.

Pereze Regadera, J.J., Gallardo, J.M., Ceballos, E.G. and Garcia, J.C.E. (1994) Model development for determination of final preferenda in freshwater species application in tench (*Tinca tinca*) *Polish Arch. Hyd.*. **42**, 27-34.

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

Richter, B.D., Braun, D.P., Mendelson, M.A. and Master, L.L. (1997) Threats to imperiled freshwater fauna, *Cons. Biol.* **11**, 1081-1093.

Rowe, D.K. (2004) Potential effects of tench (Tinca tinca) in New Zealand freshwater ecosystems, NIWA project: BOP04221.

Vainikka, (2003) Tench, *Tinca tinca* L. www.cc.jyu.fi/~ansvain/suutari/index.html Yilmaz, F. (2002) Reproductive biology of the tench *(Tinca tinca*) (L., 1758) inhabiting Porsuk Dam Lake (Kutahya, Turkey). *Fish. Res.* **55**, 313-317.

Witten, J.H., and Frank, E. (2000) *Data mining: practical machine learning tools and techniques with Java implementations*, San Francisco: Morgan Kaufman publishers p.369.

(Received: Nov.-2009, Accepted: May. 20-2010)