Ecological Informatics 4 (2009) 234–242

Contents lists available at ScienceDirect

Ecological Informatics

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

Use of different approaches to model presence/absence of *Salmo marmoratus* in Piedmont (Northwestern Italy)

Tina Tirelli [a,*], Luca Pozzi [b], Daniela Pessani [a]

[a] Dipartimento di Biologia Animale e dell'Uomo, Via Accademia Albertina 13-10123 Torino, Italy
[b] New York University, Department of Anthropology, 25 Waverly Place, New York, NY10003, USA

ARTICLE INFO

Article history:
Received 16 March 2009
Received in revised form 10 July 2009
Accepted 13 July 2009

Keywords:
Discriminant function analysis
Logistic regression
Decision tree
Artificial neural network
Sensitivity analysis
Species prediction

ABSTRACT

In Piedmont (Italy) the environmental changes due to human impact have had profound effects on rivers and their inhabitants. Thus, it is necessary to develop practical tools providing accurate ecological assessments of river and species conditions. We focus our attention on *Salmo marmoratus*, an endangered salmonid which is characteristic of the Po river system in Italy. In order to contribute to the management of the species, four different approaches were used to assess its presence: discriminant function analysis, logistic regression, decision tree models and artificial neural networks. Either all the 20 environmental variables measured in the field or the 7 coming from feature selection were used to classify sites as positive or negative for *S. marmoratus*. The performances of the different models were compared. Discriminant function analysis, logistic regression, and decision tree models (unpruned and pruned) had relatively high percentages of correctly classified instances. Although neither tree-pruning technique improved the reliability of the models significantly, they did reduce the tree complexity and hence increased the clarity of the models. The artificial neural network (ANN) approach, especially the model built with the 7 inputs coming from feature selection, showed better performance than all the others. The relative contribution of each independent variable to this model was determined by using the sensitivity analysis technique. Our findings proved that the ANNs were more effective than the other classification techniques. Moreover, ANNs achieved their high potentials when they were applied in models used to make decisions regarding river and conservation management.

1. Introduction

Globally, freshwaters are rapidly deteriorating and hence these systems are receiving increasing attention (Allan and Flecker, 1993; Matson et al., 1997; Postel, 2000). In Italy, and especially in Piedmont, there has been considerable impact of human activities on rivers. Nutrient balances have been altered both with agricultural run-offs and urban sewage discharges. Sediment inputs have increased through a combination of deforestation, floods, and road building. These changes have had profound effects on rivers and their inhabitants. Thus, there is a need for the development of practical tools providing accurate ecological assessments of river and species conditions, ultimately in order to develop measures allowing habitat and species preservation. Moreover, we need to find out in depth the relationship between the environment and the occurrence of the organisms inhabiting rivers and streams. This is fundamental for conservation management and river restoration. To reach these goals, modeling is becoming a more and more important tool for perfecting decision-making and management policies.

Freshwater modeling has made substantial progress over the last decade. Still, these ecosystems are very complex and hence hard to understand despite the substantial improvements made in ecosystem modeling and computation (Recknagel, 2002) and despite the development of highly reliable models.

Over the last several years, researchers have been applying machine learning methods to ecology more and more (Lek and Guégan, 1999; Debeljak et al., 2001; Recknagel, 2001; Dzeroski and Todorovski, 2003; Dakou et al., 2007; Goethals et al., 2007; Lencioni et al., 2007; Pivard et al., 2008). In fact, ecosystems characteristically show highly complex nonlinear relationships among their input variables. Thus machine learning techniques offer several advantages over traditional statistical analysis. Principally, they introduce fewer prior assumptions about the relationships among the variables. There are many machine learning techniques that could be applied, but decision trees (Quinlan, 1986), artificial neural networks (Lek and Guégan, 1999), fuzzy logic (Barros et al., 2000), and Bayesian belief networks (Adriaenssens et al., 2004) are seemingly the most effective for habitat suitability modeling, as has been demonstrated (Goethals and De Pauw, 2001; Dakou et al., 2007).

In the present study we focus our attention on the marble trout *Salmo marmoratus* (Cuvier, 1817), an endangered salmonid that can be distinguished from other *Salmo* species on the basis of its color pattern

* Corresponding author. Tel.: +39 011 6704538; fax: +39 011 6704508.
E-mail address: santina.tirelli@unito.it (T. Tirelli).

1574-9541/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2009.07.003

and morphology (Delling, 2002). Its range is restricted to southern Switzerland, northern Italy and the Adriatic basin of Slovenia, Croatia, Montenegro and Albania. In Italy, it is characteristic of the Po River system. The primal home range includes the left feeders of the Po River, the Tanaro River and some direct tributary basins of the northern Adriatic Sea. This home range is slowly but inexorably contracting. The upper courses of Alpine rivers and streams, which represent its typical habitat, nowadays show very few specimens belonging to this species. The decreasing presence of *S. marmoratus* is mainly due to repeated stocking with fish-farm stocks originating from the Atlantic side (*Salmo trutta fario*), causing hybridization (Gandolfi et al., 1991; Giuffra et al., 1996), due to water drawing, due to the alteration of river bottoms, and due to pollution.

Since the populations started to decline in the early 1900s and are presently still threatened, *S. marmoratus* is listed in the Annex II of the European Union Habitats Directive 92/43/CEE and in the Red List (IUCN, 1996).

Piedmont Region started some years ago to conduct fish fauna management, with the main aim of restoring, where possible, suitable habitats for the autochthonous fish communities. For this reason, the Piedmont Region passed a Regional Law (L.R. number 37 dated 29/12/06), which lays down new regulations specifically for the management of aquatic fauna, habitat and fishing. In particular, this law provides for a series of activities aimed at re-establishing consistent populations of native species. The first and most important species that needs to be conserved is *S. marmoratus*.

The present study aims at evaluating the reliability of various current classification techniques in modeling the presence/absence of *S. marmoratus* and at comparing how these techniques perform in relationship to each other. We used varieties of multivariate statistics and of the machine learning approach — in the former case, discriminant function analysis and logistic regression and, in the latter, classification and regression trees and artificial neural networks.

Of the machine learning techniques, artificial neural networks have been used quite often in ecological modeling during the last 15 years (e.g.: Lek et al., 1996; Lek and Guégan, 1999; Scardi, 1996, 2001; Mastrorillo et al., 1997; Paruelo and Tomasel, 1997; Recknagel et al., 1997; Manel et al., 1999a, 2001; Tourenq et al., 1999; Maier and Dandy, 2000; Brosse et al., 2001; Reyjol et al., 2001; Olden, 2000, 2003; Olden and Jackson, 2001, 2002; Olden et al., 2002; Joy and Death, 2004; Oakes et al., 2005). In contrast, decision trees have been used sporadically (see Dakou et al., 2007). All four techniques were applied on a dataset of 198 samples collected in different sites in Piedmont.

Such models do not require researchers to have knowledge in detail of the properties of the studied system, but generally they do require that researchers have some knowledge in order to then formulate the model and some data in order to calibrate it. Researchers must already have some previous information on the system's behavior in order to derive such a model. On the other hand, ANNs do not require researchers to have any *a priori* knowledge of the underlying system itself. Yet, they are computational tools that can represent complex nonlinear systems. ANNs combine nonlinear functions of inputs in order to model a determined output. The combination of functions is optimized by training the network in order to best match the output of the network with the desired value (Haykin, 1999). The ANN approach mimics the synaptic processes in the brain. ANNs are able to adapt themselves dynamically to highly complex problems, reproducing the dynamic interaction of multiple factors simultaneously. Researchers have applied ANNs in many fields of study recently — medicine (Chesnokov, 2008; Grossi and Buscema, 2007), ice-condition forecasts (Wang et al., 2008), species-presence prediction (Tirelli and Pessani, in press), animal-movement paths (Dalziel et al., 2008), modelization of product–user preferences (Mas et al., 2008), and predictions of motor vehicle crashes (Xie et al., 2007).

2. Materials and methods

2.1. Study area and data collection

The study system consisted of 198 sites (Fig. 1) located along the rivers in Piedmont, covering a total area of 25.399 km^2.

Salmo marmoratus was recorded at 67 of the sampling sites, corresponding to 33.38%.

Because data mining approaches are data-driven, they present researchers with a key problem — which input variables to choose in order to build the model. When large numbers of inputs are used in data mining approaches, the models become more complex, the calculation times increase, the field data collection efforts increase, and the models become less clear. In this research project, we chose variables according to their importance for fish fauna, as testified to in the literature and by expert knowledge (Lek et al., 1996; Mastrorillo et al., 1997; Olden and Jackson, 2001; Reyjol et al., 2001; Joy and Death, 2002, 2004; Laffaille et al., 2003; Olden et al., 2006). Therefore we considered the following data set: 1) altitude; 2) length of the sampling area; 3) homogeneity in width of the sampled tract (classes 0–5; the larger the widths of the sections examined, the larger the value); 4) amount of human impact (classes 0–5; the larger the impact, the larger the value); 5) amount of shade (classes 0–5; the larger the shade, the larger the value); 6) shelters for fish, visually assessed as the area consisting of undercut banks, macrophyte cover and debris jams (classes 0–5; the larger the cover, the larger the value); 7) percentage of bottom vegetation (algae and macrophytae) (classes 0–5; the larger the vegetation, the larger the value); 8–9–10) percentages (values from 0 to 100%, not classes of percentage) of the sampled area with waterfalls classified according to their heights — 8) falls with heights >1 m, 9) 0.5 m $\leq$ high ≤ 1 m, and 10) <0.5 m; 11–12–13) percentages (0–100%, not classes) of the sampled area classified according to water speed and depth — 11) riffles (areas of quite fast water with a broken-surface appearance), 12) pools (areas of slow, quite deep water with a smooth surface appearance), and 13) flat reaches (areas with smooth constant depth and water speed), each reach surveyed and estimated visually; 14–19) percentages (0–100%, not classes) of the sampled area classified according to ground surface — 14) bedrock, 15) boulders and pebbles, 16) medium gravel (≥ 1 cm), 17) little gravel (1 cm $<$ dimensions ≤ 2 mm), 18) sand (dimensions <2 mm) and 19) silt; and 20) pH.

We chose to include the 'length of the sampling area' as a predictive variable. Although hard to handle, the 'length of the sampling area' is an important variable. The longer the sampling area, the fewer the false negatives and the more certain the sampling. The weight of this variable is undoubtedly higher in the samples where *S. marmoratus* was found or was present but not detected.

The presence/absence of fish data as well as the values of all 20 variables in each site was obtained from the "Monitoraggio della fauna ittica in Piemonte" (Piedmont Region, 2006). The data were collected by a team of skilled ichthyologists in the spring, summer and fall of 2004. They used two types of single-pass electrofishing 1) with a battery-powered electric fishing machine (A.G.K. IG 200/2) operated at 150–300 V (the voltage varying according to the water conductivity); and 2) with an internal-combustion-engine machine (EFKO FEG. 8000), used when the water was deeper than 1.5–2 m.

2.2. Data-set preprocessing

We normalized the data proportionally before we used a data set to build different models. We normalized so that, normalizing between 0 and 1, minimum and maximum of all river- and habitat-measured data ranged between 0.05 and 0.95. In addition, we selected attributes through the application of 2 different techniques — the best-first search (Witten and Eibe, 2005) and Goldberg's genetic algorithm (1989) (D'heygere et al., 2006; Obach et al., 2001; Schleiter et al.,

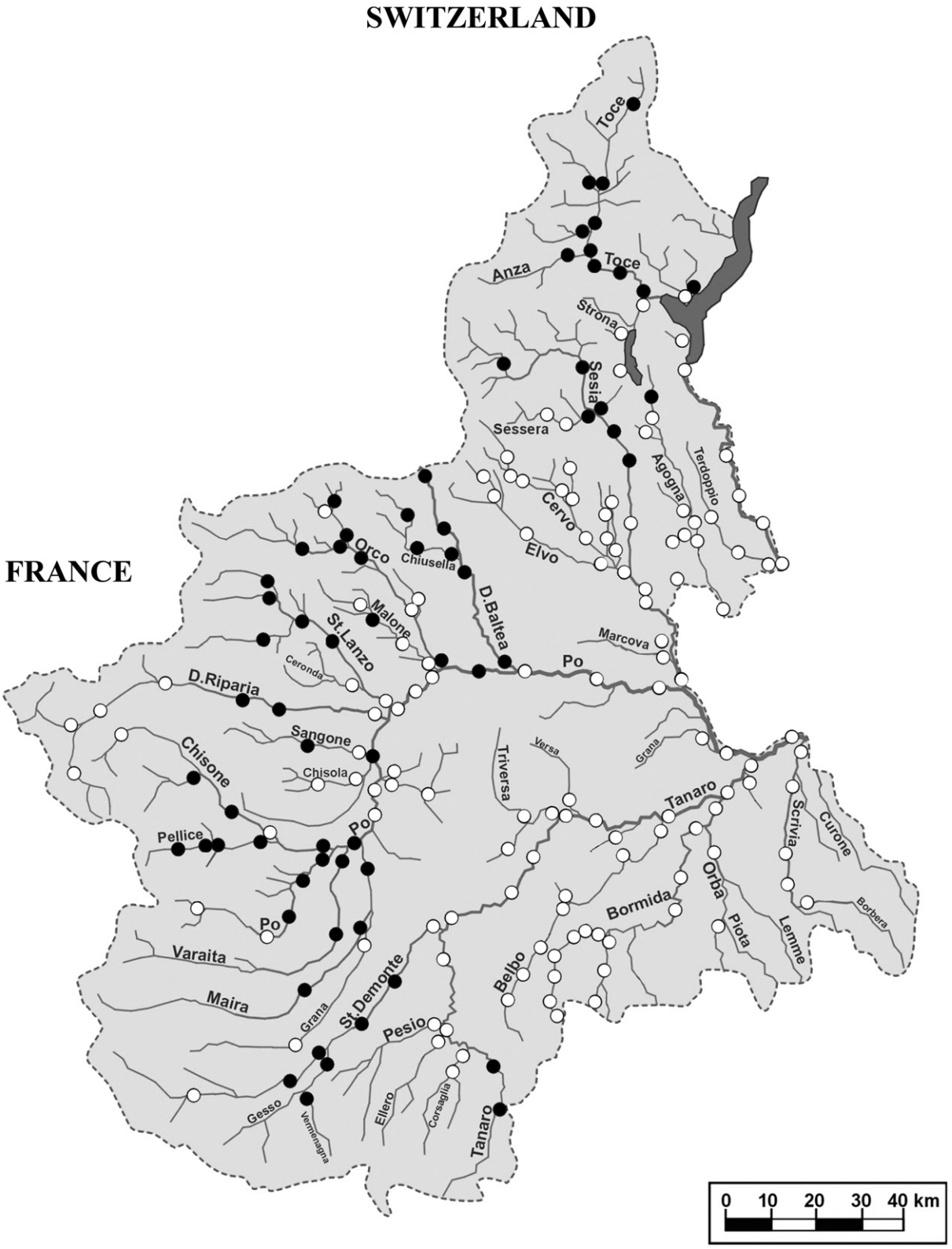

Fig. 1. Map of the Piedmont Region showing the distribution of the 198 sampling sites: in black, positive samplings for *Salmo marmoratus*, in white, negative samplings.

2001; Goldberg, 1989). These techniques involved searching among the attributes for the subsets most likely to predict the class and ultimately helped us obtain a subset of 7 inputs. These subsets are altitude, length of the sampling area, shelters for fish, boulders and pebbles, little gravel, silt, and pH.

We used discriminant function analysis, logistic regression, decision trees and artificial neural networks for the classification phase, including both the initial set of 20 features and the subset features resulting from the feature selection.

2.3. Discriminant function analysis and logistic regression classification

We performed DFA and LR, in order to distinguish between sites inhabited by marble trout and sites where they were not present. We

used species presence or absence as the dependent variable and we used river and habitat variables as the independent variables. Both the analyses were performed with a forward stepwise entry of independent variables.

The performances either of DFA or LR were estimated from a leave-one-out jack-knifing involving a holdout procedure repeated 10 times using a model derived from a calibration set of 80% of the sites — a model which, in turn, was applied to the remaining test sites.

We performed the Mann-Whitney test in order to compare the performance of the models before and after feature selection.

2.4. Decision tree models

2.4.1. Model building

We aimed to induce rules in the form of decision trees. Hence we used the "*top-down induction of decision trees*", a common technique (Quinlan, 1986), in order to generate rules relating the values of inputs with the presence/absence of marble trout.

We used the J48 algorithm in our experiments both with and without a binary split. Briefly reviewing, the J48 is the Java re-implementation of the C4.5 algorithm (Quinlan, 1993), one of the most well known and widely used decision tree induction methods. A binary split is a parameter of the J48 algorithm that decides whether a node can split into just two branches or into more than two branches. We ran the non-parametric Mann-Whitney test in order to compare the performances of the models built with binary splits and those built with multiple splits. The outputs of the models were discrete variables (presence or absence of *Salmo marmoratus*), but all the inputs were continuous.

2.4.2. Optimization techniques

We applied the tree-pruning optimization method to reduce the effects of noise in the data and to improve performance with regard to the complexity and accuracy of the predictions. Tree-pruning is a common way to cope with tree complexity. Optimal tree-pruning eliminates errors due to data noise and therefore reduces the size of models and makes them clearer and more accurate in their classification (Bratko, 1989). We chose post-pruning rather than forward-pruning. We used both of the post-pruning operations available — subtree replacement and subtree raising — and then compared the results obtained. We controlled the intensity of pruning by changing the confidence factor — a parameter affecting the error rate estimate in each node between 0.15 and 0.25. The smaller the confidence factor, the larger the difference between the error rate estimates of a parent node and its potential splits. Thus, the smaller the confidence factor, the larger the possibility that splits are replaced by leaves.

2.4.3. Model validation

We aimed to assess model performances. Hence we evaluated five parameters on the basis of matrices of confusion (Fielding and Bell, 1997): 1) the percentage of Correctly Classified Instances (CCI), frequently used when presence/absence of taxa is predicted; 2) model sensitivity (the ability to predict species presence accurately); 3) model specificity (the ability of the model to correctly predict species absences); 4) Cohen's kappa coefficient (Cohen, 1960); and 5) the area under the receiver-operating-characteristic (ROC) curve.

The CCI are affected by the frequency of occurrence of the test organism(s) being modeled (Fielding and Bell, 1997; Manel et al., 1999b). Thus Cohen's Kappa coefficient is a more reliable performance measure of presence/absence models because Cohen's k is negligibly affected by prevalence (e.g. Dedecker et al., 2004, 2005; D'heygere et al., 2006). Cohen's k gives a rather conservative estimate of prediction accuracy because it underestimates agreements due to chance (Foody, 1992). However, k values come from the information content of the dataset, which has limited extractable information. For this reason, Gabriels et al. (2007) suggest that different disciplines may show differences in k threshold values. They assess the following k values in a freshwater ecological context: 0.00–0.20: poor; 0.20–0.40: fair; 0.40–0.60: moderate; 0.60–0.80: substantial; and 0.80–1.00: excellent. In the area under the ROC curve, a value of 0.7 indicates satisfactory discrimination, a value of 0.8 good discrimination and 0.9 very good discrimination (Hosmer and Lemeshow, 2000).

Model training and validation were based on stratified 10-fold cross-validation (Kohavi, 1995). In order to estimate a reliable error of the models, 10-fold cross-validation experiments were repeated 10 times and the average predictive performance was calculated. This experimental procedure allowed us to determine the 95% confidence limit of the average predictive performance. We ran Mann-Whitney test in order to compare the performances of the models built respectively before and after feature selection (unpruned and pruned).

2.5. Artificial neural network models

2.5.1. Model architecture

We built three layered feed-forward neural networks with bias and developed them with the following architecture. Namely, there were two options: 1) networks built using the initial set of 20 features and 2) networks built using the subset features resulting from the feature selection. In each case there was 1 output node — for the presence/absence of *Salmo marmoratus*. There was 1 hidden layer between the input and the output layers with a number of neurons optimized by trial and error. A single layer shortens the computation time and often yields the same result as ANNs with multiple hidden layers (Kurkova, 1992; Bishop, 1995). The number of hidden neurons was chosen to minimize the trade-off between network bias and variance (Bishop, 1995). We determined the optimal number of hidden neurons empirically, by comparing the performances of different networks. We tested many different types of architecture — with different values of learning rates and momentums (range 0.1–0.5), different numbers of epochs and different numbers of neurons in the hidden layer — until we obtained the best predicting model. Then we chose the simpler models — those with fewer hidden nodes — from the range of those with similar performances. We did this for two reasons. Firstly, the simpler network of two with similar performances is likely to predict new cases better (Bishop, 1995). Secondly, the optimal network geometry corresponds to the smallest network that captures the relationships in the training data adequately (D'heygere et al., 2006).

2.5.2. Model optimization

We used cross-validation to be sure not to overtrain the networks. Without cross-validation, in fact, a network can overfit the training data and so be rendered unable to generalize previously unseen data during the test phase. To this end, k-fold cross-validated neural networks were trained using the error-back-propagation algorithm (Rumelhart et al., 1986). The learning rate was set to 0.3 and the momentum set to 0.2 for the initial set of 20 features. Both learning rate and momentum were set to 0.2 for the subset features resulting from the feature selection. Cross-validation method is particularly useful when the number of cases is limited. During k-fold cross-validation, the data set is equally split into k parts, the ANN model is trained with k-1 parts, and validated with the rest. The procedure is repeated k times. Goethals et al. (2007) suggest building different models using a set of combinations of k between 3 and 10, of k = number of cases/2 and k = number of cases − 1, in order to determine the best k value. Low k values can build robust ANN models, which, however, usually have relatively low performances. For this reason, a high k value should be used when there are just a few cases in the dataset. Because of this fact, the 'leave-one-out' cross-validation method (Efron, 1983) is often used in ecological study (Guégan et al., 1998; Brosse et al., 2001, 2003; Beauchard et al., 2003). For these reasons, we determined the optimal k value empirically by comparing the performances of different cross-validated networks where k = 3–10, 99, 197. Then, we performed the Mann-Whitney tests to compare the statistical differences. The comparison was done on the basis of five

parameters (those described in Materials and methods/Decision tree models/Model validation). We then chose the model with the lowest k value from among models with similar performances.

2.5.3. Model validation

We followed the same procedure as for DT models except for the fact that we ran 10 stratified k-fold cross-validation experiments in order to estimate the reliable error of the models. Then, we calculated the average predictive performances and chose one of the 10 networks (either among the 20-input networks or the networks with the 7 features resulting from the feature selection) included within this range as the final network for our model. This experimental procedure enabled us to determine a 95% confidence limit of the average predictive performance. We performed the Mann-Whitney test to compare the performance of the ANN models built with different number of inputs and with decision tree models.

2.5.4. Model interpretation

We aimed to determine the importance of the inputs on the outputs. Hence we chose sensitivity analysis from among the different techniques available, as suggested by Hunter et al. (2000). A sensitivity analysis indicates which input variables are considered the most important by that particular neural network. Hunter et al. (2000) analyzed sensitivity by replacing each variable in turn with random values and assessing the effect of this upon the output error (i.e. the RMS of the individual cross-entropy errors of the test cases). A variable that is relatively important will cause a correspondingly large deterioration in the model's performance. The more sensitive the network is to a particular input, the greater the deterioration we can expect, and therefore the greater the ratio. Once sensitivities have been calculated for all variables, they may be ranked in order.

3. Results

3.1. Discriminant function analysis and logistic regression classification

Table 1 shows the mean performances of DFA and LR. The performances were calculated from a leave-one-out jack-knifing involving a holdout procedure repeated 10 times. We used a model derived from a calibration set of 80% of the sites and then applied this model to the remaining test sites. The mean performances and standard deviations of DFA and LR were calculated on models built using both all 20 river and habitat variables and only the 7 variables coming from feature selection. We detected significant differences in sensitivity and specificity between DFA and LR models. According to the tests, DFA had significantly higher sensitivity values than LR ($p<0.05$). Conversely, LR had significantly higher specificity values than DFA ($p<0.05$). This resulted both in models with 20 inputs and those with 7 inputs.

3.2. Decision tree models

We induced models for the prediction of the environment suitability of marble trout by using the J48 algorithm with binary and multiple splits.

The average and standard deviations of the five performance parameters were calculated either for DT with a binary split or for DT with a multiple split built using all the 20 features (Table 2). These parameters were 1) percentage CCI, 2) sensitivity, 3) specificity, 4) Cohen's k statistic, and 5) the area under the ROC curve of the 10 repeated 10-fold cross-validated unpruned and pruned DTs before feature selection.

We performed the Mann-Whitney tests to compare all of the five parameters used to assess for the performance among the multiple models, among the binary models, and between multiple and binary models before feature selection. No significant differences in the predictive performance in any of the above tests were detected. Therefore, for the trees built using the 7 inputs coming from feature selection, we used only binary splits, on the basis of the paper by Dakou et al. (2007), who obtained positive results in a similar freshwater context. The average and standard deviations of the five performance parameters for these last models were calculated (Table 3). The percentage of CCI was not always very high. Cohen's k statistic was relatively high in all the cases, but not enough to consider the models reliable. In fact, these values of Cohen's k revealed that most of the predictions were based on chance. Sensitivity always reached high values (>78.6%), while specificity was around 60.0%. The area under the ROC curve (0.73) indicated satisfactory discrimination.

Tree-pruning was performed for DTs built both before and after feature selection (Tables 2 and 3). The unpruned trees had included a large number of leaves, which made them more complex and hindered an ecological interpretation (Tables 2 and 3). Pruning usually allows researchers to make the models less complex and the performance more efficient. Thus models with different intensities of post-pruning were induced by varying the confidence factor between 0.15 and 0.25. The optimal confidence factor was 0.15 for the trees built before feature selection and 0.17 for the trees built after feature selection.

No significant increases were detected in predictive performances among unpruned and pruned trees built using the 7 inputs. No significant increases were detected also in predictive performances between DTs built using 20 and 7 inputs.

3.4. ANN models

The optimization of the number of hidden neurons by trial and error resulted in two different network architectures, according to the number of inputs used.

In models built before feature selection, the final architecture showed 20 input neurons, 10 hidden ones and 1 output. The performances and reliabilities of the ANNs did not improve with $k>10$ among the different k-fold cross-validations that were tested for these models. In fact, there were no statistical differences between 10- and 99-fold cross-validated ANNs, according to the results of the Mann-Whitney tests performed on the five parameters assessing the performances of the ANNs. Thus we used 10-fold cross-validation to build our model.

We calculated the average and standard deviation of percentage CCI, of sensitivity, of specificity, of Cohen's k statistic and of the area under the ROC curve of the 10 repeated 10-fold cross-validated ANNs

Table 1
Performance of discriminant function analysis (DFA) and logistic regression (LR) executed before and after feature selection, for predicting the presence/absence of *Salmo marmoratus*.[a]

	CCI				Sen				Spe			
Model	DFA20	DFA7	LR20	LR7	DFA20	DFA7	LR20	LR7	DFA20	DFA7	LR20	LR7
Mean	75.70	74.78	75.18	74.93	74.00	72.91	49.66	48.21	76.64	75.74	88.95	89.17
s.d.	2.13	3.04	3.26	3.51	4.22	4.80	9.37	10.71	3.26	2.81	3.91	3.54

Performances were calculated from a leave-one-out jack-knifing involving a holdout procedure repeated 10 times.

[a] Percentage of correctly classified instances = CCI; sensitivity = sen; specificity = spe; discriminant function analysis using 20 inputs = DFA20; discriminant function analysis using 7 inputs = DFA7; logistic regression using 20 inputs = LR20; logistic regression using 7 inputs = LR7; standard deviation = s.d.

Table 2
Predictive results of decision tree models based on the J48 algorithm without pruning and with post-pruning optimization.[a]

	Decision tree		CCI	*k*	sen	spe	ROC
Binary split	un	Mean	72.23	0.38	79.17	58.79	0.71
		s.d.	2.54	0.06	2.61	4.68	0.02
	subre	Mean	72.23	0.37	79.17	59.28	0.71
		s.d.	2.54	0.05	2.94	3.97	0.03
	subra	Mean	72.80	0.39	79.51	60.00	0.73
		s.d.	1.92	3.94	2.48	33.82	0.03
Multiple split	un	Mean	72.23	0.38	79.18	58.79	71.19
		s.d.	2.41	0.05	2.48	4.44	2.34
	subre	Mean	71.74	0.37	78.27	59.29	70.99
		s.d.	2.22	0.05	2.94	3.97	2.83
	subra	Mean	72.80	0.39	79.51	60.00	72.63
		s.d.	1.92	0.04	2.48	3.38	2.58

The models were built with 20 inputs.
[a] Unpruned = un; subtree replacement = subre; subtree raising = subra; percentage of correctly classified instances = CCI; sensitivity = sen; specificity = spe; Cohen's k = k; area under the ROC curve = ROC; standard deviation = s.d.

(Table 4). We chose the final network from among these 10 inputs. This network had quite a high percentage of CCI (78.31%), of sensitivity (87.03%), of specificity (61.43%), a moderate k coefficient (0.50) and an area under the ROC curve of 0.77. This last indicates that the model discriminates in a satisfactory way (Hosmer and Lemeshow, 2000).

In models built after feature selection, the final architecture showed 7 input neurons, 4 hidden ones and 1 output. The performances and reliabilities of the ANNs did not improve with $k>5$ among the different k-fold cross-validations that were tested for the 7 input models. For example, we report the results of the Mann-Whitney test performed on the five parameters assessing for the comparative performances of the ANNs in 5- and in 10-fold cross-validation (10 is the most common in use). The tests highlighted statistical differences only in k values, showing that the ANNs with 5-fold cross-validation performed significantly better than 10-fold cross-validation ones ($p<0.05$). Thus we used 5-fold cross-validation to build our model.

The average and standard deviations were calculated of percentage CCI, of sensitivity, of specificity, of Cohen's k statistic and of the area under the ROC curve of the 10 repeated 5-fold cross-validated ANNs (Table 4). The final network was chosen among these 10 inputs. The final network showed very good performance and accuracy. This network had a high percentage of CCI (81.82), a very high sensitivity (86.48%) and a quite good specificity (74.83%). It had a very good k coefficient (0.60) (Gabriels et al., 2007), one that yielded substantial model performance. It had an area under the ROC curve of 0.85, one that indicated that the model discriminated very well (Hosmer and Lemeshow, 2000).

In addition, we conducted the Mann-Whitney tests to compare the performances of the ANNs built with 20 and 7 features, respectively

Table 3
Predictive results of decision tree models based on the J48 algorithm without pruning and with post-pruning optimization.[a]

Decision trees — binary splits		CCI	*k*	sen	spe	ROC	nl	cl
un	Mean	71.26	0.35	78.63	60.41	0.73	19	–
	s.d.	2.51	0.05	0.02	0.05	0.03		
subre	Mean	71.51	0.36	79.07	59.64	0.73	11	0.17
	s.d.	2.10	0.04	0.01	0.03	0.03		
subra	Mean	71.46	0.36	78.99	59.69	0.73	11	0.17
	s.d.	2.51	0.04	0.01	0.03	0.03		

The models were built using the 7 inputs coming from feature selection.
[a] Unpruned = un; subtree replacement = subre; subtree raising = subra; percentage of correctly classified instances = CCI; sensitivity = sen; specificity = spe; Cohen's k = k; area under the ROC curve = ROC; number of leaves = nl; confidence level = cl; standard deviation = s.d.

Table 4
Predictive results of ANNs before and after feature selection.[a]

ANN		CCI	*k*	sen	spe	ROC
20 inputs	Mean	73.09	0.41	79.55	60.52	0.78
	s.d.	3.40	0.07	4.84	5.61	0.02
7 inputs	Mean	76.46	0.47	82.23	66.54	0.81
	s.d.	3.44	0.08	2.79	5.88	0.03
Mann-Whitney test	*p*	<0.05	<0.05	n.s.	<0.05	<0.05

[a] Percentage of correctly classified instances = CCI; sensitivity = sen; specificity = spe; Cohen's k = k; area under the ROC curve = ROC; standard deviation = s.d.; significance = p.

(Table 4), showing that the ANNs built using 7 inputs performed better than the ANNs built using 20 inputs ($p<0.05$).

Finally, we conducted the Mann-Whitney tests to compare the performances of the 7-feature ANNs and the 7-feature binary DTs. The tests showed that the ANNs performed better than the DTs ($p<0.05$) in all cases.

The final ANN model we obtained clearly showed the influence of all the 7 inputs used in modeling the presence/absence of the species in Piedmont. The sensitivity analysis allowed us to rank the input neurons according to their importance in building the model in the following way: altitude, little gravel, boulders and pebbles, length of the sampling area, silt, pH, and shelters for fish.

4. Discussion and conclusion

The objective of this study was to assess the effectiveness and to compare the performances and reliabilities of different approaches to predicting *Salmo marmoratus* presence in the Piedmont Region, approaches that potentially can be applied more widely. In particular, we compared multivariate techniques and machine learning methods. Over the last decade, machine learning methods have become more and more popular for modeling ecological ecosystems (e.g.: Lek and Guégan, 1999; Debeljak et al., 2001; Recknagel, 2001; Dzeroski and Todorovski, 2003; Brewer et al., 2007; Dakou et al., 2007; deKock and Wolmarans, 2007a,b; Goethals et al., 2007; Lencioni et al., 2007; Pivard et al., 2008). Their popularity is mainly due to the advantages they offer over traditional statistical analyses when applied to ecosystems. Natural systems, in fact, are characterized by highly complex nonlinear relationships. Among machine learning techniques, ANNs and decision trees have been recently shown to have a high potential in habitat suitability modeling (Goethals and De Pauw, 2001; Dakou et al., 2007). For these reasons, we decided to use these tools together with discriminant function analysis and logistic regression in order to model the presence of an endangered autochthonous species (*S. marmoratus*) and compare the results obtained. The aim is to contribute to *S. marmoratus* management by finding the approach that performs better and better describes the relationships of species with the habitats they occupy.

The results we obtained are highly encouraging. In fact, these results show that we can predict marble trout presence in Piedmont with reasonable accuracy. They show how useful it is to compare different approaches to modeling Alpine freshwater fish. In addition, we can apply these techniques in other Italian regions and even in other countries reasonably confidently.

4.1. Discriminant function analysis and logistic regression classification

These multivariate techniques allowed us to classify approximately 75% of the sites correctly. In general DFA has higher sensitivity and lower specificity values than LR both in models with 20 and those with 7 inputs. No differences in performance were detected between DFA models with 20 and 7 inputs. There were not any differences between LR models with 20 and 7 inputs, too. This is due to the fact that both

the techniques use only variables of interest and not uninformative ones.

Not all the modeling procedures performed well. This clashed with the findings of Manel et al. (1999a). The performances of the LR models we obtained are similar to those obtained by Manel et al. (1999a). Nevertheless, the ANNs outperformed not only the LR models but also all the other models. Our LR models showed much lower sensitivity than specificity, as did those of Manel et al. (1999a), in spite of the occurrence of *S. marmoratus* at 33.38% of the sampled sites. Here, the prevalence effect warrants careful consideration.

4.2. Decision tree models

The decision tree models did not perform very well in each set of models when the percentage of CCI and specificity were used as evaluation measures. In fact, the CCI was ca. 71–72% and the specificity was ca. 59–60%. Moreover, Cohen's *k* statistic showed that the models yielded unreliable predictions, in that most of the classifications were based on chance. On the contrary, the decision tree models did perform well in relation to sensitivity (above 78%) and the area under the ROC curve (satisfactory discrimination).

It is not clear why the models cannot make reliable predictions despite the fact that the inputs we used do seem to be valid indicators of various features of habitat suitability. Yet, these same inputs did produce good classifications when used to run the other kinds of models (DFA, LR, ANN). This is an important point.

Decision tree models may not be among the best performing techniques to apply for predicting presence/absence of *Salmo marmoratus*. In fact, they do not perform well in relation to other species and even taxa. For example, decision tree models did not perform well in predicting macroinvertebrate taxa (Dakou et al., 2007). Their Cohen's *k* values were even lower than those in the present study.

DT models were too complex, especially the unpruned trees with their many leaves. Consequently, the results of DT models prevent an ecologically interpetation. In fact, the trees could not yield any information about the habitat suitability of *S. marmoratus*. The J48 algorithm produced very detailed trees, hampering the models' ability to generalize. Therefore, we used pruning optimization to reduce tree complexity. This allowed us to obtain simpler trees that could be interpreted ecologically. We used just the post-pruning technique to improve the performance and reduce the complexity of the models because post-pruning was shown to perform the best (Dakou et al., 2007). Tree-pruning did not result in a significant improvement of any of the five parameters used to assess for model performances. Pruning does decrease the complexity of the trees and the variance. However, it increases the bias and hence only improves the accuracy of a model slightly (Geurts, 2000; Dakou et al., 2007).

The DT models may not have performed well because of the dimensions of the dataset and because of the fact that the frequency of occurrence of *S. marmoratus* was lower than 50%. In fact, the predictive performance of models based on decision trees is strongly related to the frequency of occurrence of the predicted taxa (Goethals et al., 2001; Manel et al., 2001).

We used tree-pruning to gain an improvement in the clarity of the models that are induced. This is the most important goal when models are built to make decisions about river restoration and conservation management.

Moreover, feature selection allowed us to reduce the size of the problem but did not help the DT models improve their performances. In decision trees, the J48 algorithm chooses the most suitable attribute to split at each branch of a tree. Therefore, we may expect that the less suitable attributes would be chosen out. However, things are quite different in practice. One researcher (John et al., 1994) reported that performance decreased around 5–10% after a random binary attribute had been added as an extra variable in a dataset. This was a feature that we preferred to test. Therefore we can state, in this specific case, that DTs can handle large-dimension datasets. The most evident side effect of the use of a large number of inputs is an increase in computational cost. Therefore, even though feature selection does not affect the DTs performance, we recommend using feature selection in order to contain computational costs.

4.3. ANN models

This approach is to be considered the most effective of the four tools that we analyzed, a tool that contributes to the management and the conservation of the species. In fact, ANN models allow us 1) to understand the factors contributing to the presence/absence of the species, using a reliable and modern technique and 2) to incorporate multiple input parameters into a single model in a fast and flexible way.

The ANN approach performed better than DFA, LR and DT. In particular, the present research shows that the accuracy of ANN classification improves when a small set of optimally selected features is used. This is due to the improvement of signal-to-noise ratio and to the reduction of overfitting. The selection stage serves to eliminate all but the most relevant attributes and thus allows us to reduce the number of input variables. We performed the selection in order to help the models predict more accurately (D'heygere et al., 2003, 2006; Tirelli and Pessani, in press).

Moreover, the findings of the present research make it evident that learning in ANNs is sensitive to the input data used. When researchers select the appropriate features through preprocessing, the performances of their models in an ecological context are improved considerably. Preprocessing calls for just a small additional effort since preprocessing techniques are not time/computing intensive. This is true for any learning algorithm since the complexity of the data used directly affects the learning algorithm's performance (Piramuthu, 2004). In fact, the irrelevant information in ANNs without variable selection passes through the nodes, possibly influences the connection weights slightly, and affects the overall performance of ANNs. Moreover, variable selection allows researchers to decrease ANN size. This reduces computational costs, increases speed, and uses less data to estimate connection weights efficiently.

When proper feature selection is applied in ecological contexts, it makes a crucial contribution to species management and to the planning of protective measures, especially in regard to presence/absence predictions of endangered taxa.

The high performance level of ANN modeling makes it the most useful technique for application to *Salmo marmoratus* management. *S. marmoratus* were present at 33.38% of the sites. For this reason, the 10 ANN models predicted the high average number of presences of the species correctly and the final network that was chosen was highly sensitive. On the other hand, low occurrence causes low sensitivity. This coincides with the results of other research projects (Manel et al., 1999a, 2000, 2001; Olden and Jackson, 2002; Olden et al., 2002; Oakes et al., 2005). In general, ANNs, like DTs, predict more accurately when there are more occurrences of the species under examination (Spitz et al., 1996; Mastrorillo et al., 1997; Manel et al., 1999a; Tourenq et al. (1999). ANNs predict especially accurately when the number of presences and absences is just about the same (Tourenq et al., 1999). This is obviously a problem because in ecology, especially for rare species, absences are of course more frequent than presences. High levels of correctly predicted presence are particularly important in instances where the presence of scarce species is predicted for conservation purposes — for example, in identifying areas for protection or management of rare species. This fact underlines once more the value of the ANN approach to modeling *S. marmoratus* presence. Moreover, the ability of ANNs to classify previously unseen cases will eventually allow researchers to use the network reported here to analyze new inputs coming from different geographic areas in order to plan larger scale protection actions. ANN technique allows for

an objective, statistically robust, site-specific prediction of the species presence to be made, without extensive field analysis, therefore permitting researchers to save time and money during the survey phase of the project. This is due to the fact that the model used is reliable even though based on few, easily detectable, environmental parameters.

We should emphasize something. In our research project, we built both DTs and ANNs accurately — varying a series of parameters for each kind of approach — in order to obtain the most reliable and best performing models (Olden, 2007). Secondly, we ran 10 stratified *k*-fold cross-validation experiments and then calculated the average predictive performance. In this way we found the range of performances of the models. We also ran 10 experiments in the case of DFA and LR and then calculated the average predictive performance. Out of all these models, it was the 10 ANN experiments that outperformed the other experiments.

The performances of the reduced model that we obtained also suggest that there was an accurate link between *S. marmoratus* and the variables used to build the model itself. This again underlies the fact that researchers need to perform a correct feature selection before running the models. In addition, the sensitivity analysis showed that all the 7 input variables considered had effects on the presence/absence of *S. marmoratus*. This further demonstrated the correctness of the input choices that we made.

After this discussion on data processing, our findings do have something to report about the physical environment of *S. marmoratus* itself — i.e. elevation, bottom granulometry, fine suspended sediments, and shelters.

Altitude plays a great role in building the models because it is good integrator of the thermal conditions. In this regard, our findings coincide with findings on *Salmo trutta* (Baran et al., 1993, 1995; Lek et al., 1996) and with findings on the fish community inhabiting the Wellington Region (Joy and Death, 2004).

Bottom reaches of boulders and pebbles are important for the presence of marble trout, as our findings seem to show. In this regard, these findings on the granulometry of the bottom are only partially in agreement with studies on brown trout (Vlach et al., 2005; Scheurer et al., 2009). In fact, Vlach et al. (2005) have often found *Salmo trutta* adult specimens on sandy or muddy bottom. In contrast, we found that bottom reaches of boulders and pebbles were important for the presence of marble trout. Seemingly, marble trout avoid sand and silt bottoms because they need to avoid physical alteration and infections or because these bottoms interfere with their reproductive behavior.

Fine suspended sediments are an important factor for *S. marmoratus*. Prolonged exposure to fine suspended sediments can affect fish health and behavior (Alabaster and Lloyd, 1980; Newcombe and MacDonald, 1991; Berry et al., 2003; Scheurer et al., 2009). Specimens show signs of sub-lethal stress at values of fine sediments lower than 90 mg/l. These symptoms include changes in blood chemistry, gill- or skin- epithelia damage, and the resulting increased number of infections. An environment with fine sediments can even cause higher mortality rates (Berry et al., 2003; Scheurer et al., 2009). Exposure to fine sediments during the incubation period can delay the emergence of fry, alter the natural emergence pattern (Fudge et al., 2008; Scheurer et al., 2009), and have negative repercussions on the entire reproductive cycle. Also, fine sediment reduces interstitial flow and oxygen supply, increasing embryo mortality and decreasing emergence success (Chapman, 1988).

Another important factor for *S. marmoratus* is the number of shelters. This species thrives when there are more shelters. This coincides with findings that shelters have a great impact on brown trout density when they reach percentages lower than 2% (Lek et al., 1996).

Our research project leads us to conclude in this way. We recommend that researchers use various techniques to build presence/absence models in order to compare their prediction accuracy and their performances and in order to evaluate the importance of the input variables used to build the model. Such a use of different approaches allows researchers to avoid the risk of choosing a model that is not properly suited to approaching the problem, which is what would have happened if the present research project had been based on fewer approaches.

Acknowledgments

We would like to thank the Progetto Lagrange of Fondazione Cassa di Risparmio di Torino for funding this project.

References

Adriaenssens, V., Goethals, P.L.M., Charles, J., De Pauw, N., 2004. Application of Bayesian Belief Networks for the prediction of macroinvertebrate taxa in rivers. Annales of Limnology — International Journal of Limnology 40 (3), 181–191.
Alabaster, J.S., Lloyd, R., 1980. Water Quality Criteria for Freshwater Fish. Food and Agriculture Organization of the United Nations, Butterworths, London. 297 pp.
Allan, J.D., Flecker, A.S., 1993. Biodiversity conservation in running waters. BioScience 43, 32–43.
Baran, P., Delacoste, M., Lascaux, J.M., Belaud, A., 1993. Relations entre les caractéristiques de l'habitat et les populations de truites communes (*Salmo trutta* L.) de la vallé e de la Neste d'Aure. Bulletin française de la Pêche et de la Piscicolture 331, 321–340.
Baran, P., Delacoste, M., Dauba, F., Lascaux, J.M., Belaud, A., 1995. Effects of reduced flow on brown trout (*Salmo trutta* L.) populations downstream dams in French Pyrenees. Regulated Rivers: Research and Management 10, 347–361.
Barros, L.C., Bassanezi, R.C., Tonelli, P.A., 2000. Fuzzy modelling in population dynamics. Ecological Modelling 128, 27–33.
Beauchard, O., Gagneur, J., Brosse, S., 2003. Macroinvertebrate richness patterns in North African streams. Journal of Biogeography 30, 1821–1833.
Berry, W., Rubinstein, N., Melzia, B., Hill, B., 2003. The biological effects of suspended and bedded sediment (SABS) in aquatic systems: a review. USEPA, Washington D.C. 58 pp.
Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, United Kingdom. 504 pp.
Bratko, I., 1989. Machine learning. In: Gilhooly, K.J. (Ed.), Human and Machine Problem Solving. Pelnum Press, New York and London, pp. 265–287.
Brewer, S.K., Rabeni, C.F., Sowa, S.P., Annis, G., 2007. Natural landscape and stream segment attributes influencing the distribution and relative abundance of riverine smallmouth bass in Missouri. North American Journal of Fisheries Management 27 (1), 326–341.
Brosse, S., Lek, S., Townsend, C.R., 2001. Abundance, diversity, and structure of freshwater invertebrates and fish communities: an artificial neural network approach. New Zealand Journal of Marine and Freshwater Research 35, 135–145.
Brosse, S., Arbuckle, C.J., Townsend, C.R., 2003. Habitat scale and biodiversity: influence of catchment, stream reach and bedform scales on local invertebrate diversity. Biodiversity Conservation 12, 2057–2075.
Chapman, D.W., 1988. Critical-review of variables used to define effects of fines in redds of large salmonids. Transactions of the American Fisheries Society 117, 1–21.
Chesnokov, Y., 2008. Complexity and spectral analysis of the heart rate variability dynamics for distant prediction of paroxysmal atrial fibrillation with artificial intelligence methods. Artificial Intelligence in Medicine 43, 151–165.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1), 37–46.
Cuvier, G., 1817. Le Regne Animal, distribue d'apres son organisation, pour servir de base a l'histoire naturelle des animaux et d'introduction a l'anatomie comparee. Deterville, Paris. 653 pp.
D'heygere, T., Goethals, P.L.M., De Pauw, N., 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. Ecological Modelling 160 (3), 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 (1–2), 20–29.
Dakou, E., D'heygere, T., Dedecker, A.P., Goethals, P.L.M., Lazaridou-Dimitriadou, M., De Pauw, N., 2007. Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). Aquatic Ecology 41, 399–411.
Dalziel, B.D., Morales, J.M., Fryxell, J.M., 2008. Fitting probability distributions to animal movement trajectories: using artificial neural networks to link distance, resources, and memory. American Naturalist 172, 248–258.
Debeljak, M., Dzeroski, S., Jerina, K., Kobler, A., Adamic, M., 2001. Habitat suitability modelling for red deer (*Cervus elaphus* L.) in South-central Slovenia with classification trees. Ecological Modelling 138, 321–330.
Dedecker, A., Goethals, P.L.M., Gabriels, W., De Pauw, N., 2004. Optimisation of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium). Ecological Modelling 174 (1–2), 161–173.
Dedecker, A.P., Goethals, P.L.M., De Pauw, N., 2005. Sensitivity and robustness of stream model based on artificial neural networks for the simulation of different management scenarios. In: Lek, S., Scardi, M., Verdonschot, P.F.M., Descy, J.P., Park, Y.S. (Eds.), Modelling Community Structure in Freshwater Ecosystems. Springer-Verlag, pp. 133–146.

deKock, K.N., Wolmarans, C.T., 2007a. Distribution and habitats of *Corbicula fluminalis africana* (Mollusca: Bivalvia) in South Africa. Water South Africa 33 (5), 709–715.
deKock, K.N., Wolmarans, C.T., 2007b. Distribution and habitats of the alien invader freshwater snail *Physa acuta* in South Africa. Water South Africa 33 (5), 717–722.
Delling, B., 2002. Morphological distinction of the marble trout, *Salmo marmoratus*, in comparison to marbled *Salmo trutta* from River Otra, Norway. Cybium 26, 283–300.
Dzeroski, S., Todorovski, L., 2003. Learning population dynamics models from data and domain knowledge. Ecological Modelling 170, 129–140.
Efron, B., 1983. Estimating the error rate of a prediction rule: some improvements on cross-validation. Journal of the American Statistical Association 78, 316–331.
Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38–49.
Foody, G.M., 1992. On the compensation for chance agreement in image classification accuracy assessment. Photogrammetric Engineering and Remote Sensing 58, 1459–1460.
Fudge, T.S., Wautier, K.G., Evans, R.E., 2008. Effect of different levels of fine-sediment loading on the escapement success of rainbow trout fry from artificial redds. North American Journal of Fisheries Management 28 (3), 758–765.
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 networks. Aquatic Ecology 41, 427–441.
Gandolfi, G., Zerunian, S., Torricelli, P., Marconato, A., 1991. I pesci delle acque interne italiane. Istituto Poligrafico e Zecca dello Stato, Roma. 617 pp.
Geurts, P., 2000. Some enhancements of decision tree bagging. In: Zighed, D.A., Komorowski, J., Zytkow, J. (Eds.), Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge discovery. Springer-Verlag, Berlin, pp. 136–147.
Giuffra, E., Guyomard, R., Forneris, G., 1996. Phylogenetic relationships and introgression patterns between incipient parapatric species of Italian brown trout (*Salmo trutta* L. complex). Molecular Ecology 5, 202–207.
Goethals, P., De Pauw, N., 2001. Development of a concept for integrated ecological assessment in Flanders, Belgium. Journal of Limnology 60, 7–16.
Goethals, P.L.M., Džeroski, S., Vanrolleghem, P., De Pauw, N., 2001. Prediction of benthic macro-invertebrate taxa (Asellidae and Tubificidae) in watercourses of Flanders by means of classification trees. IWA 2nd World water congress, Berlin, pp. 5–6.
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 Algorithm in Search. Addison-Winsley Publishing Company, Reading, Optimization and Machine Learning.
Grossi, E., Buscema, M., 2007. Introduction to artificial neural networks. European Journal of Gastroenterology & Hepatology 19, 1046–1054.
Guégan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384.
Haykin, S., 1999. Neural Networks: a Comprehensive Foundation. Upper Saddle River, Prentice Hall, NJ. 842 pp.
Hosmer, D., Lemeshow, S., 2000. Applied Logistic RegressionSecond Edition. A. Wiley-Interscience Publication, John Wiley and Sons Inc, New York, NY. 373 pp.
Hunter, A., Kennedy, L., Henry, J., Ferguson, I., 2000. Application of neural networks and sensitivity analysis to improved prediction of trauma survival. Computer Methods and Programs in Biomedicine 62, 11–19.
IUCN, 1996. *1*996 IUCN Red List of Threatened Animals. IUCN, Gland, Switzerland.
John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant Features and the Subset Selection Problem. Proceedings of the 11th International Conference on Machine Learning ICML94, 121–129.
Joy, M.K., Death, R.G., 2002. Predictive modelling of freshwater fish as a biomonitoring tool in New Zealand. Freshwater Biology 47, 2261–2275.
Joy, M.K., Death, R.G., 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology 49, 1036–1052.
Kohavi, R., 1995. A study of cross-validation and bootstrap for estimation and model selection. In: Mellish, C.S. (Ed.), Proccedings of the 14th International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publisher, Montreal, pp. 1137–1143.
Kurkova, V., 1992. Kolmogorov's theorem and multilayer neural networks. Neural Networks 5, 501–506.
Laffaille, P., Feunteun, E., Baisez, A., Robinet, T., Acou, A., Legault, A., Lek, S., 2003. Spatial organisation of European eel (*Anguilla anguilla* L.) in a small catchment. Ecology of Freshwater Fish 12, 254–264.
Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modeling, an introduction. Ecological Modelling 120, 65–73.
Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996. Role of some environmental variables in trout abundance models using neural networks. Aquatic Living Resources 9, 23–29.
Lencioni, V., Maiolini, B., Marziali, L., Lek, S., Rossaro, B., 2007. Macroinvertebrate assemblages in glacial stream systems: a comparison of linear multivariate methods with artificial neural networks. Ecological Modelling 203, 119–131.
Maier, H.G., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software 15, 101–124.
Manel, S., Dias, J.M., Ormerod, S.J., 1999a. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalaian river bird. Ecological Modelling 120, 337–347.
Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999b. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology 36, 734–747.
Manel, S., Buckton, S.T., Ormerod, S.J., 2000. Testing large-scale hypotheses using surveys: the effects of land use on the habitats, invertebrates and birds of Himalayan rivers. Journal of Applied Ecology 37, 756–770.
Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence/absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921–931.
Mas, J.A.D., Cuesta, S.A., Marzal, J.A., 2008. Application of neural networks to the modelization of products user preferences. Dyna 83, 148–156.
Mastrorillo, S., Lek, S., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biology 38, 237–246.
Matson, P.A., Parton, W.J., Power, A.G., Swift, M.J., 1997. Agricultural intensification and ecosystem properties. Science 277, 504–508.
Newcombe, C.P., MacDonald, D.D., 1991. Effects of suspended sediments on aquatic ecosystems. North American Journal of Fisheries Management 11, 72–82.
Oakes, R.M., Gido, K.B., Falke, J.A., Olden, J.D., Brock, B.L., 2005. Modelling of stream fishes in the Great Plains, USA. Ecology of Freshwater Fish 14, 361–374.
Obach, M., Wagner, R., Werner, H., Schmidt, H.H., 2001. Modelling population dynamics of aquatic insects with artificial neural networks. Ecological Modelling 146, 207–217.
Olden, J.D., 2000. An artificial neural network approach for studying phytoplankton succession. Hydrobiology 436, 131–143.
Olden, J.D., 2003. A species-specific approach to modeling biological communities and its potential for conservation. Conservation Biology 17, 854–863.
Olden, J.D., 2007. Critical threshold effects of benthiscape structure on stream herbivore movement. Philosophical Transactions of the Royal Society 362, 461–472.
Olden, J.D., Jackson, D.A., 2001. Fish-habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. Transactions of the American Fisheries Society 130, 878–897.
Olden, J.D., Jackson, D.A., 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology 47, 1976–1995.
Olden, J.D., Jackson, D.A., Peres-Neto, P.R., 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. Transactions of the American Fisheries Society 131, 329–336.
Olden, J.D., Joy, M.K., Death, R., 2006. Rediscovering the species in community-wide predictive modeling. Ecological Applications 16 (4), 1449–1460.
Paruelo, J.M., Tomasel, F., 1997. Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. Ecological Modelling 98, 173–186.
Piramuthu, S., 2004. Computing, Artificial Intelligence and Information Technology. Evaluating feature selection methods for learning in data mining applications. European Journal of Operational Research 156, 483–494.
Pivard, S., Demšar, D., Lecomte, J., Debeljak, M., Džeroski, S., 2008. Characterizing the presence of oilseed rape feral populations on field margins using machine learning. Ecological Modelling 212, 147–154.
Postel, S.L., 2000. Entering an era of water scarcity: the challenges ahead. Ecological Applications 10, 941–948.
Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.
Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA. 302 pp.
Recknagel, F., 2001. Application of machine learning to ecological modelling. Ecological Modelling 146 (1–3), 303–310.
Recknagel, F. (Ed.), 2002. Ecological informatics: understanding ecology by biologically-inspired computation. Springer-Verlag, Berlin. 432 pp.
Recknagel, F., French, M., Harkonen, P., Yabunaka, K., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecological Modelling 96, 11–28.
Reyjol, Y., Lim, P., Belaud, A., Lek, S., 2001. Modelling of microhabitat used by fish in natural and regulated flows in the river Garonne (France). Ecological Modelling 146, 131–142.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagation errors. Nature 323, 533–536.
Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Marine Ecology Progress Series 139, 289–299.
Scardi, M., 2001. Advances in neural network modeling of phytoplankton primary production. Ecological Modelling 146, 33–46.
Scheurer, K., Alewell, C., Bänninger, D., Burkhardt-Holm, P., 2009. Climate and land-use changes affecting river sediment and brown trout in alpine countries — a review. Environmental Science and Pollution Research 16, 232–242.
Schleiter, I.M., Obach, M., Borchardt, D., Werner, H., 2001. Bioindication of chemical and hydromorphological habitat characteristics with benthic macro-invertebrates based on artificial neural networks. Aquatic Ecology 35, 147–158.
Spitz, F., Lek, S., Dimopoulos, I., 1996. Neural network models to predict penetration of wild boar into cultivated fields. Journal of Biological Systems 4, 433–444.
Tirelli, T., Pessani, D., in press. Use of decision tree and artificial neural network approaches to model presence/absence of *Telestes muticellus*, in Piedmont (North-Western Italy). River Research and Applications, Published Online. doi:10.1002/rra.1199.
Tourenq, C., Aulagnier, S., Mesléard, F., Durieux, L., Gonzalez, G., Lek, S., 1999. Use of artificial neural networks for predicting rice crop damage by greater flamingos in the Camargue, France. Ecological Modelling 120, 349–358.
Vlach, P., Duŝek, J., Švátora, M., Moravec, P., 2005. Fish assemblage structure, habitat and microhabitat preference of five fish species in a small stream. Folia Zoologica 54, 421–431.
Wang, T., Yang, K.L., Guo, Y.X., 2008. Application of artificial neural networks to forecasting ice conditions of the Yellow River in the Inner Mongolia reach. Journal of Hydrologic Engeneering 13, 811–816.
Witten, I.H., Eibe, F., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann, San Francisco, CA. 525 pp.
Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. Accident Analysis and Prevention 39, 922–933.