

Modelling habitat requirements of white-clawed crayfish (*Austropotamobius pallipes*) using support vector machines

L. Favaro⁽¹⁾, T. Tirelli⁽¹⁾, D. Pessani⁽¹⁾

Received December 26, 2010

Revised March 19, 2011

Accepted April 21, 2011

ABSTRACT

Key-words:
crayfish,
machine
learning,
ecological
modelling,
conservation,
endangered
species

The white-clawed crayfish's habitat has been profoundly modified in Piedmont (NW Italy) due to environmental changes caused by human impact. Consequently, native populations have decreased markedly. In this research project, support vector machines were tested as possible tools for evaluating the ecological factors that determine the presence of white-clawed crayfish. A system of 175 sites was investigated, 98 of which recorded the presence of *Austropotamobius pallipes*. At each site 27 physical-chemical, environmental and climatic variables were measured according to their importance to *A. pallipes*. Various feature selection methods were employed. These yielded three subsets of variables that helped build three different types of models: (1) models with no variable selection; (2) models built by applying Goldberg's genetic algorithm after variable selection; (3) models built by using a combination of four supervised-filter evaluators after variable selection. These different model types helped us realise how important it was to select the right features if we wanted to build support vector machines that perform as well as possible. In addition, support vector machines have a high potential for predicting indigenous crayfish occurrence, according to our findings. Therefore, they are valuable tools for freshwater management, tools that may prove to be much more promising than traditional and other machine-learning techniques.

RÉSUMÉ

Conditions requises pour la modélisation de l'habitat de l'écrevisse à pattes blanches (*Austropotamobius pallipes*) par machines à vecteur de supports

Mots-clés :
écrevisse,
apprentissage
par la machine,
modélisation
écologique,
conservation,
espèces
en danger

L'habitat de l'écrevisse à pattes blanches a été profondément modifié dans le Piémont (nord-ouest de l'Italie) par les changements environnementaux dus aux impacts humains. En conséquence, les populations indigènes ont considérablement diminué. Dans ce projet de recherche, des machines à vecteur de supports ont été testées comme outils possibles pour évaluer les facteurs écologiques qui déterminent la présence de l'écrevisse à pattes blanches. Un ensemble de 175 sites ont été échantillonnés, dont 98 avec présence d'*Austropotamobius pallipes*. À chaque site 27 variables physico-chimiques, environnementales et climatiques ont été mesurées. Différentes méthodes de sélection ont été employées. Elles aboutissent à trois sous-ensembles de variables qui permettent de construire trois différents types de modèles : (1) des modèles sans sélection de variables; (2) des modèles construits en appliquant l'algorithme de Goldberg après sélection de variables; (3) des modèles construits en utilisant une combinaison d'estimateurs

(1) Dipartimento di Biologia Animale e dell'Uomo, Università degli Studi di Torino, Torino, Italy, livio.favaro@unito.it

gérés par filtres après sélection de variables. Ces différents modèles nous ont aidés à prendre conscience de l'importance de la bonne méthode de sélection si on veut construire des machines à vecteur de supports qui fonctionnent aussi bien que possible. De plus, les machines à vecteur de supports sont très performantes pour prédire l'occurrence des écrevisses indigènes, selon nos résultats. Par conséquent, elles sont un outil valable pour la gestion des eaux douces, outil qui semble être plus prometteur que les techniques traditionnelles et d'autres par apprentissage de machine.

INTRODUCTION

Austropotamobius pallipes (Lereboullet, 1858) is the only indigenous crayfish species living in north-western Italy. It plays an important role in regulating the biodiversity of streams and lakes (Gherardi *et al.*, 2001) and actively contributes to the flow of energy and to the cycling of matter (Nyström, 1999; Nyström *et al.*, 1999). Because of its ecological importance, the white-clawed crayfish is broadly considered as a keystone species in preserving the well-being of freshwater ecosystems (Holdich, 2003). In the past, *A. pallipes* were widely distributed along brooks and small tributaries flowing into Piedmont's main rivers. Over the last few decades, the number of sites inhabited by the species has decreased markedly, as has the density of the populations (Nardi *et al.*, 2004; Tirelli *et al.*, 2008; Favaro *et al.*, 2010). This decline has been caused mainly by (1) the pollution of water bodies due to agriculture and urban activities, (2) erosion, (3) siltation, and (4) habitat loss and fragmentation (Grandjean *et al.*, 2000; Broquet *et al.*, 2002; Favaro *et al.*, 2010). There is another threat to *A. pallipes* survival – the introduction of non-indigenous crayfish species that dramatically contribute to the spread of crayfish plague (*Aphanomyces astacii*).

Because of this alarming situation in Piedmont, *A. pallipes* is protected by a law of Piedmont Region (L.R. No. 32 of 2 November 1982). Moreover, *A. pallipes* has been included on the Red List of threatened animal species of the International Union for the Conservation of Nature and Natural Resources (Baillie and Groombridge, 1996) and in Annexes II and V of the Habitat Directive (Council of the European Communities, 1992, 1997). Therefore, several research projects have been conducted in Italy and elsewhere in Europe over the last few years to preserve the extant populations of *A. pallipes* and their sites (Holdich and Rogers, 1997; Nardi *et al.*, 2006; Tirelli *et al.*, 2008).

There are promising tools that can help us solve such environmental challenges as the loss of biodiversity – the tools that ecological informatics equips us with (Green *et al.*, 2005). Ecological informatics can be seen as an interdisciplinary framework that uses advanced computational technology to study ecological processes and patterns on various levels of ecosystem complexity (Recknagel, 2003). There is a rapidly growing area of ecological informatics called *machine learning* (ML), a tool that identifies structures in complex, nonlinear data and generates accurate predictive models.

Different applications of ML methods have been used in ecology (e.g., Fielding, 1999; Recknagel, 2001, 2003; Cushing and Wilson, 2005; Ferrier and Guisan, 2006; Park and Chon, 2007). They have proved to be powerful alternatives to traditional modelling approaches. ML methods consist of a range of approaches, including (1) artificial neural networks (Lek *et al.*, 1996; Hoang *et al.*, 2001; Dedecker *et al.*, 2007; Goethals *et al.*, 2007; Tirelli and Pessani, 2009); (2) classification and regression trees (De'ath and Fabricius, 2000; Dzeroski *et al.*, 2000; Goethals *et al.*, 2001; Dakou *et al.*, 2007; Lencioni *et al.*, 2007; Pivard *et al.*, 2008; Tirelli and Pessani, 2009; Hoang *et al.*, 2010); (3) fuzzy logic (Salski and Sperlbaum, 1991; Adriaenssens *et al.*, 2004a; Mouton *et al.*, 2009); (4) genetic algorithms and programming (Stockwell and Noble, 1992); (5) Bayesian belief networks (Adriaenssens *et al.*, 2004b); and (6) support vector machines (Vapnik, 1995; Guo *et al.*, 2005; Hu and Davis, 2005; Drake *et al.*, 2006; Shan *et al.*, 2006; Sanchez-Hernandez *et al.*, 2007a, 2007b; Ribeiro and Torgo, 2008; Hoang *et al.*, 2010).

These methods are being used more and more because they can model the complex, nonlinear relationships that typify ecological data. However, they do not have to satisfy the restrictive assumptions of conventional, parametric approaches (Guisan and Zimmermann, 2000; Peterson and Vieglais, 2001; Olden and Jackson, 2002; Elith *et al.*, 2006). Further, they allow researchers to develop highly reliable models (Recknagel, 2003).

SUPPORT VECTOR MACHINES

SVMs consist of a new group of learning algorithms. Originally developed by Vapnik (1995), they are examples of the new ML methods available. SVMs present a challenge for modellers because they are models that are statistically-based and because they guarantee performance in a theoretical way (Cristianini and Scholkopf, 2002).

They are inductive modelling techniques inspired by some features of biological information processing. They are based on an algorithm that finds the maximum-margin hyperplane – *i.e.* the hyperplane showing the greatest separation between the classes. The instances at the minimum distances from the maximum-margin hyperplane constitute the support vectors. There is at least one support vector for each class, but there is often more than one support vector. The maximum-margin hyperplane for the learning problem is something that is defined exclusively by the set of support vectors. Once the support vectors are found, the maximum-margin hyperplane is easy to build. Statistically, the optimal boundaries should be generalised to unseen cases with the least errors among all possible boundaries separating the classes, therefore reducing the confusion among classes as much as possible. The support vectors are placed on the very edge of the class distributions inside the border region separating the classes. Therefore, they are elements that are critical for the training set. All the other training instances are irrelevant to the extent that they can be omitted without changing either the position or orientation of the hyperplane (Witten and Frank, 2005; Sanchez-Hernandez *et al.*, 2007a). These support vectors are placed closest to the decision boundary. Therefore, the classifier uses extreme cases to separate the two classes from each other. Detailed descriptions of SVMs are available in Vapnik (1995) as well as Tax and Duin (2004). SVMs are equipped with several features that are more advantageous than those of other ML techniques:

- (1) overfitting is unlikely to occur (Vapnik, 1995; Burges, 1998; Duda *et al.*, 2001). This is due to the fact that overfitting is associated with instability while the maximum margin hyperplane is relatively stable, moving only if training instances corresponding to SVs are added or deleted.
- (2) SVMs can obtain subtle and complex decision boundaries and hence produce results that are more competitive than those of the best current accessible-classification methods.
- (3) SVMs yield excellent generalisation performance while tackling a wide range of problems, particularly while solving numerous nonlinear regression and time-series problems, as evidenced over the years by Vapnik (1995), Hoang *et al.* (2010) and others.
- (4) SVMs require only a minimum of model tuning because only a few parameter settings need to be adjusted (Decoste and Scholkopf, 2002; Guo *et al.*, 2005; Hoang *et al.*, 2010).
- (5) Only a small training dataset is required to find the optimal separating hyperplane (Sanchez-Hernandez *et al.*, 2007a). This is very important in the ecological applications of the method because researchers can reach the levels of performance and the accuracy of models built using other ML techniques. Thus, researchers can work in a more elegant way, sampling only a much smaller number of sites of extreme spectral response (Sanchez-Hernandez *et al.*, 2007a). They can sample only the locations that are important for training the classifier, thus avoiding conventional, more expensive sampling approaches (Sanchez-Hernandez *et al.*, 2007a).

On the other hand, there are disadvantages of SVMs: they are computationally complex and they are slow. Even so, because of their many advantages, SVMs have been applied successfully to many tasks. Nevertheless, they have only very recently been applied to ecological predictions (Joachims, 1998; Brown *et al.*, 2000; Cristianini and Scholkopf, 2002; Decoste

and Scholkopf, 2002; Huang *et al.*, 2002; Guo *et al.*, 2005; Hu and Davis, 2005; Shan *et al.*, 2006; Sanchez-Hernandez *et al.*, 2007a, 2007b; Ribeiro and Torgo, 2008; Hoang *et al.*, 2010). Because of their novelty and their potential usefulness in ecological applications, we decided to build models of *A. pallipes* presence using the SVM approach. Our goal consisted of assessing the reliability of the models and testing their performance in a freshwater context.

The aim of the present research project is two-fold:

(1) To use SVMs to model the species' presence in Piedmont; to compare the SVM performance with the performances of logistic regression, decision trees and artificial neural networks (Tirelli *et al.*, submitted); to ascertain whether SVMs are reliable modelling techniques for investigating the habitat requirements of white-clawed crayfish;

(2) To compare the performance of SVMs built under two different sets of circumstances: (a) those built without performing feature selection and (b) those built with the use of only those features that stem from a previous feature selection procedure. This comparison was made because feature reduction is an open question. In fact, some authors (Hoang *et al.*, 2010) deem feature reduction unnecessary for SVM classification, while others advocate feature reduction in order to make the classification and performance of the models more accurate (Sanchez-Hernandez *et al.*, 2007a, 2007b). Therefore, we aimed at understanding whether researchers do or do not need feature selection for the specific task of white-clawed crayfish modelling.

MATERIALS AND METHODS

> STUDY AREA AND DATA COLLECTION

Samplings were made in 175 sites across Piedmont, covering a total area of 25399 km² (Fig. 1). The presence of *A. pallipes* was recorded in 98 sites. These sites are located within brooks and small tributaries flowing into the Po River. Most of them were certainly inhabited by native crayfish until a few decades ago (Gelder *et al.*, 1999). The geological substrate ranges from siliceous to calcareous. Therefore, the physical and chemical characteristics also range in features across the 175 sites.

Species presence was assessed during the daylight hours through manual surveys (2 people for 1 h). At night traps were used (50 × 25 × 25 cm with a 3-mm mesh size, baited with pig or chicken liver, left overnight). Samplings were performed from late spring to early autumn each year from 2005 to 2009. Samplings were always performed during normal regimen flow, hence never after heavy rains or during dry spells.

> INPUT VARIABLE CHOICE

Because data-mining approaches are data-driven, they present researchers with the key problem of choosing which input variables they need for building the model. In this research project, variables were chosen according to their importance for *A. pallipes* presence, as outlined by several authors (see, e.g., Broquet *et al.*, 2002; Trouilhé *et al.*, 2007; Brusconi *et al.*, 2008; Favaro *et al.*, 2010).

> ENVIRONMENTAL VARIABLES

We measured the following environmental variables on site: altitude; width at moderate flow; width at high flow; percentages (0–100%, not classes) of the sampled area classified according to granulometry – bedrock (fixed rock), boulders and pebbles (> 3 cm), medium gravel (> 1 cm), little gravel (1 cm < dimension < 2 mm), sand and silt (dimension < 2 mm); water velocity; amount of shade (classes 0–5; the larger the shade, the larger the value).

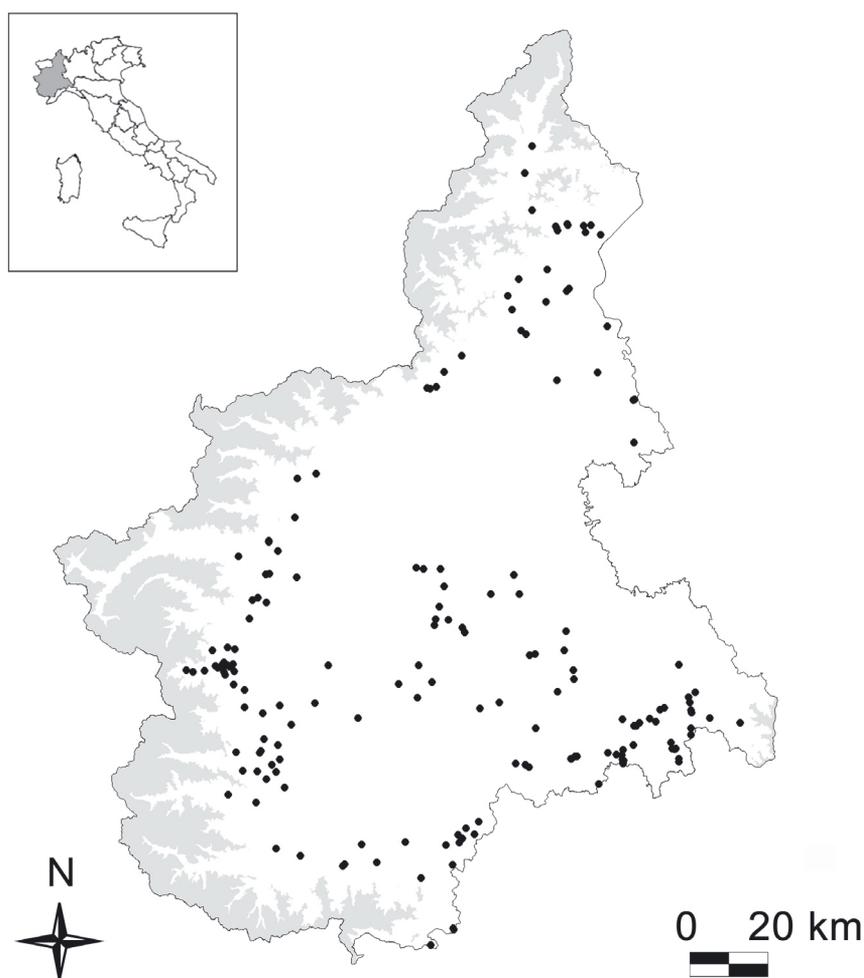


Figure 1
Sampling sites in Piedmont, NW Italy.

Figure 1
Sites échantillonnés dans le Piémont, nord-ouest de l'Italie.

> **PHYSICAL-CHEMICAL VARIABLES**

At each site, we measured pH, conductivity and the percentage of dissolved oxygen (DO) by means of a multi-parameter probe (mod. Hydrolab Quanta). Moreover, we collected two 100-mL water samples at about 15 cm depth, one that allowed us to avoid floating materials. The samples were stored in sterile polythene test tubes and frozen until chemical analysis. The concentrations of several inorganic ions commonly used to assess water quality were measured: ammonium (NH_4^+), nitrates (NO_3^-), orthophosphate (PO_4^{3-}), chlorides (Cl^-), sulphates (SO_4^{2-}), calcium (Ca^{2+}) and magnesium (Mg^{2+}). These concentrations were measured using a DR LANGE Lasa 100 spectrophotometer. The BOD_5 index was also evaluated. The measured ion concentrations were considered to be constant because we sampled the sites during normal flow regimens. We made this consideration even though ion concentrations are not generally conservative variables.

> CLIMATE VARIABLES

The software DIVA-GIS version 5.4.0.1 (www.diva-gis.org) was used with raster data taken directly from BIOCLIM. This software consists of a bioclimatic prediction system that uses surrogate terms (bioclimatic parameters) derived from mean monthly climate estimates in order to approximate energy and water balances at given locations (Nix, 1986). BIOCLIM derives its bioclimatic parameters through the use of monthly or weekly values of maximum temperature, minimum temperature, rainfall, radiation and evaporation.

In this research project, the following parameters were used: (1) the annual mean temperature (= the mean of all the weekly mean temperatures, where each weekly mean temperature is the mean of that week's maximum and minimum temperatures); (2) the maximum temperature of the warmest period (= the highest temperature of any weekly maximum temperature); (3) the minimum temperature of the coldest period (= the lowest temperature of any weekly minimum temperature); (4) the annual precipitation (= the sum of all the monthly precipitation estimates); (5) the precipitation of the wettest period (= the precipitation of the wettest month); and (6) the precipitation of the driest period (= the precipitation of the driest month).

> DATASET PRE-PROCESSING

Data was proportionally normalised before using a dataset to build the different models. Normalisation was done in such a way that the minimum and maximum of all the environmental, physical-chemical and climate data ranged between 0.05 and 0.95.

We built models both without performing feature selection (hereafter: NFS) – *i.e.* with the use of all 27 measured variables. We also built models that depended on previous feature selections – *i.e.* with the exclusive use of features that came out of a previous feature selection.

Feature selections were performed generally by searching the space of attribute subsets through combining an attribute-subset evaluator with a search method. Filter methods were applied, methods that selected features on the basis of the measures of feature predictability and redundancy. Supervised filters are very flexible and allow researchers to combine various search and evaluation methods. In particular, the five supervised-filter evaluators were combined with the Ranker search method: (1) χ^2 (Chan and Wong, 1991), (2) information gain, (3) gain ratio (Quinlan, 1990), (4) symmetrical uncertainty, and (5) OneR, available in the package WEKA (<http://www.cs.waikato.ac.nz/ml/weka>; Witten and Frank, 2005). Moreover, the Cfs Subset Evaluator (Hall, 1998) was combined with Goldberg's genetic algorithm (1989) search method. Genetic algorithms like Goldberg's are commonly used in river ecology, but they are used in combination with such classifiers as (1) classification and regression trees, and (2) artificial neural networks (Obach *et al.*, 2001; Schleiter *et al.*, 2001; D'heygere *et al.*, 2006; Tirelli *et al.*, 2009). The algorithms used in each evaluator are described thoroughly by Witten and Frank (2005). Each evaluator establishes the particular method used in order to assign a worth to each subset of features. Thus, the search method fixes the style of the performed search.

Feature selection can be done either by using the full training set or by cross-validation. We used ten-fold cross-validation for each of the five methods. These techniques involved searching among the attributes for the subsets most likely to predict the class. They yielded two subsets of inputs. We used SVMs for the classification phase, including both the initial set of 27 features and the two feature subsets resulting from the feature selection.

> MODEL DEVELOPMENT

SVMs use Platt's sequential minimization algorithm (SMO) for training a support vector classifier (Platt, 1998, 1999; Keerthi *et al.*, 2001). This implementation replaces all missing values and transforms nominal attributes into binary ones. Platt's sequential minimization algorithm is also included in the machine-learning package WEKA (<http://www.cs.waikato.ac.nz/ml/weka>;

Witten and Frank, 2005). We chose SMO because it is extremely easy to implement, because it is often faster than other algorithms, and because it has better scaling properties.

We applied the polynomial Kernel. We did not modify the default values of the parameter settings in the WEKA toolbox except for the exponents of the polynomial Kernel. Different exponents from 1.0 to 5.0 were tested to improve the performance of the SVM models (Hoang *et al.*, 2010). The model with the best-performing exponent was chosen.

For each of the two subsets from feature selection as well as for the 27 inputs, we used k -fold cross-validation. Goethals *et al.* (2007) suggest that the best k -value can be determined by building three different models: (1) models using a set of combinations of k between 3 and 10, (2) models using a set of combinations of k corresponding to the number of cases/2, and (3) models using a set of combinations of k corresponding to the number of cases – 1. For these reasons, we determined the optimal k value empirically by comparing the performances of different cross-validated SVMs using the Mann-Whitney U test.

Both for SVMs built using all the measured inputs and for SVMs built using the two feature selection subsets, the model with the best-performing exponent and the best-performing k value was run ten times after randomisation to estimate any eventual reliable error. At this point, we ran the non-parametric Mann-Whitney U test to compare the performance of the models built before feature selection and those built afterwards.

The performance of predictive models can be assessed by calculating the percentage of sites where the presence/absence of the studied taxa is predicted correctly (Manel *et al.*, 2001). However, correctly classified instances (CCI) are affected by the frequency of occurrence of the test organism(s) being modelled (Fielding and Bell, 1997; Manel *et al.*, 1999; Dedecker *et al.*, 2002). To compensate, we used the following additional performance measures to assess the models, namely: (1) model sensitivity (ability to predict species presence accurately); (2) model specificity (ability to predict species absence accurately); (3) Cohen's k coefficient (Cohen, 1960); (4) and the area under the receiver-operating-characteristic (ROC) curve. Cohen's k is a measure of the proportion of all possible cases of presence or absence that are predicted correctly after accounting for chance effects. Thus, Cohen's k interprets the predictive performance of the models better than CCI alone 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). Models with $k > 0.4$ and CCI $> 70\%$ are to be considered reliable (D'heygere *et al.*, 2006; Dakou *et al.*, 2007; Goethals *et al.*, 2007; Hoang *et al.*, 2010). Moreover, Gabriels *et al.* (2007) suggest that different disciplines may show differences in k threshold values. Hence, they assess the following k values in a freshwater ecological context too, confirming the ranges suggested by Landis and Koch (1977), which are classified as 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, almost perfect. Regarding the area under the ROC curve, a value of 0.7 indicates satisfactory discrimination, a value of 0.8 good discrimination and a value of 0.9 very good discrimination (Hosmer and Lemeshow, 2000).

RESULTS AND DISCUSSION

> FEATURE SELECTION PHASE

The first subset, resulting from using Goldberg's genetic algorithm search method (1989), is made up of the following 11 inputs: (1) PO_4^{3-} , (2) NH_4^+ , (3) BOD_5 , (4) Ca^{2+} , (5) water hardness, (6) pH, (7) conductivity, (8) % of bedrock, (9) maximum temperature of the warmest period, (10) annual precipitation, and (11) precipitation of the wettest period. Hereafter, we indicate the SVMs built using these inputs with GA.

The second subset is made up of a unique core of relevant features, resulting from using five filter evaluators combined with the Ranker search method. The selected core is made up of the features present in the first 15 positions of the rankings. They are: (1) PO_4^{3-} , (2) NH_4^+ , (3) NO_3^- , (4) Ca^{2+} , (5) BOD_5 , (6) DO percentage saturation, (7) pH, (8) conductivity, (9) % of

Table I

Results of the Mann–Whitney *U* tests performed to compare the performances of the different models (models without feature selection = NFS; models built after genetic algorithm feature selection = GA; models built after selecting inputs using the four supervised-filter evaluators = 15I; percentage of correctly classified instances = CCI; sensitivity = Sen; specificity = Spe; Cohen's k = k ; area under the ROC curve = ROC).

Tableau I

Résultats du test *U* de Mann-Whitney pour la comparaison des performances des différents modèles (modèles sans sélection = NFS ; modèles construits après sélection selon l'algorithme Genetic = GA ; modèles construits après entrées sélectionnées utilisant les quatre estimateurs gérés par filtres = 15I ; instances correctement classées (CCI), sensibilité (Sen), spécificité (Spe) et kappa de Cohen (k), et aire en dessous de la courbe ROC (ROC)).

	<i>k</i> -fold	CCI	Sen	Spe	Cohen's k	ROC
NFS	10 vs. 87	n.s.	n.s.	n.s.	n.s.	n.s.
GA	9 vs. 10	n.s.	n.s.	n.s.	n.s.	n.s.
15I	10 vs. 87	n.s.	n.s.	n.s.	n.s.	n.s.

Table II

Performances of support vector machine models built before and after feature selection (models without feature selection = NFS; models built after genetic algorithm feature selection = GA; models built after selecting inputs using the four supervised-filter evaluators = 15I; percentage of correctly classified instances = CCI; sensitivity = Sen; specificity = Spe; Cohen's k = k ; area under the ROC curve = ROC; standard deviation = s.d.).

Tableau II

Performances des machines à vecteur de supports construites avant et après sélection de variables (modèles sans sélection = NFS ; modèles construits après sélection selon l'algorithme Genetic = GA ; modèles construits après entrées sélectionnées utilisant les quatre estimateurs gérés par filtres = 15I ; instances correctement classées (CCI), sensibilité (Sen), spécificité (Spe) et kappa de Cohen (k), et aire en dessous de la courbe ROC (ROC) ; déviation standard = s.d.).

			CCI	Sen	Spe	Cohen's k	ROC
NFS	exponent = 1.1	mean	73.92	83.71	61.45	0.46	0.73
	k = 10	s.d.	1.09	0.94	2.28	0.02	0.01
GA	exponent = 2.5	mean	70.84	84.44	53.44	0.39	0.69
	k = 9	s.d.	1.69	2.22	2.47	0.03	0.02
15I	exponent = 4.1	mean	76.62	79.03	73.60	0.52	0.76
	k = 10	s.d.	2.33	2.25	3.42	0.04	0.02

bedrock, (10) water velocity, (11) amount of shade, (12) width at moderate flow, (13) altitude, (14) minimum temperature during the coldest period, and (15) precipitation during the wettest period. Hereafter, we indicate the SVMs built using these inputs with 15I.

> MODELS

The best-performing models were obtained using an exponent of 1.1 for NFS models, of 2.5 for GA models, and of 4.1 for 15I SVMs.

The optimal k value was determined empirically by comparing the performances of different cross-validated SVM models using the Mann–Whitney *U* test. The results appear in Table I. Thus, we used ten-fold cross-validation to build NFS models, nine-fold cross-validation for GA SVMs, and ten-fold cross-validation for 15I models. Table II and Figure 2 illustrate the performances of the three different models according to the inputs used to build the SVMs.

According to the CCI, k and ROC thresholds, the presence/absence of *A. pallipes* can be predicted reliably by SVMs. The average CCI for NFS and 15I SVMs was higher than the criteria for good model performance. Meanwhile, the mean CCI value for GA models was just

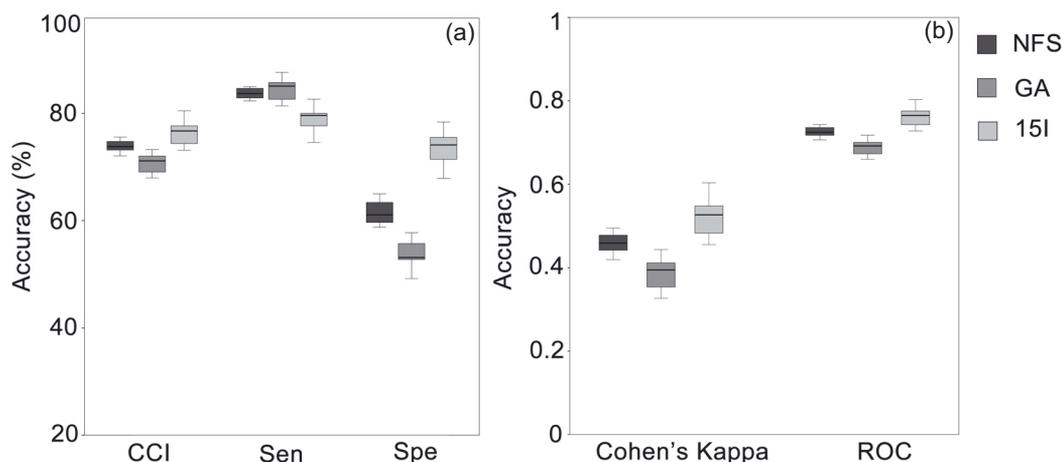


Figure 2

(a) Correctly classified instances (CCI), sensitivity (Sen), specificity (Spe); and (b) Cohen's kappa and area under the ROC curve (ROC) of SVMs obtained using the three different variable datasets (models without feature selection = NFS; models built after genetic algorithm feature selection = GA; models built after selecting inputs using the four supervised-filter evaluators = 15I).

Figure 2

(a) Instances correctement classées (CCI), sensibilité (Sen), spécificité (Spe); et (b) kappa de Cohen et aire en dessous de la courbe ROC (ROC) des SVM obtenus avec trois jeux de données différents (modèles sans sélection = NFS; modèles construits après sélection selon l'algorithme Genetic = GA; modèles construits après entrées sélectionnées utilisant les quatre estimateurs gérés par filtres = 15I).

slightly above 70%, which is the threshold-limit value for considering a model reliable. The same trend is followed for the mean value of k , which has a k lower than 0.4 in GA SVMs. At the same time, the mean k value obtained by NFS and 15I SVMs is higher than this threshold. Moreover, we conducted Mann-Whitney U tests to assess the statistical differences in the performances of the three types of models – (1) the 10 repeated ten-fold cross-validated NFS models, (2) the 10 repeated nine-fold cross-validated GA models, and (3) the 10 repeated ten-fold cross-validated 15I models. The tests showed that the NFS SVMs performed better than the GA one ($p < 0.01$), except for sensitivity. Moreover, the test showed that the best predictions were obtained with 15I SVMs ($p < 0.01$), except for sensitivity. Therefore, even with quite a small dataset, support vector machines can make predictions well.

SVMs with GA feature selection performed worse than NFS SVMs. This may be explained by the fact that selecting inputs can cause the loss of meaningful information about the impact of environmental, physical-chemical and climate variables on crayfish presence. This result absolutely confirms what Hoang *et al.* (2010) relate. Moreover, the performances of SVMs improve after feature selection, when performed using the four supervised-filter evaluators. This suggests that the choice of the proper subset of features is fundamental for SVMs too.

Indeed, although feature reduction may not be mandatory for SVM-based classification, using a proper feature selection method can increase the classification accuracy significantly, as stated by Sanchez-Hernandez *et al.* (2007a, 2007b). These models are more able to deal with a higher number of variables than other techniques such as decision trees and artificial neural networks. Even so, these models still benefit from appropriate feature selection, improving their performance.

Moreover, SVMs outperform both decision trees (DTs) and artificial neural networks (ANNs) built using the same 175-sample site-presence data by Tirelli *et al.* (submitted). Both DTs and ANNs had the same 15 inputs from the 15I feature selection. Therefore, we can assert that SVMs are the most reliable, best-performing, and most useful tools for crayfish management, better than the various techniques we had used in modelling presence/absence of *A. pallipes*

in Piedmont – better than logistic regression (Favaro *et al.*, 2010) and better than other data-mining techniques (Tirelli *et al.*, submitted).

In the first place, SVMs may very well be more suitable than DTs because SVMs can simultaneously assess the effect of all driving variables on the presence of the species. In contrast, DTs can only evaluate one variable at each branch of the tree. In the second place, SVMs produce more accurate and reliable classifiers than ANNs because SVMs can obtain more subtle and complex decision boundaries. SVMs are more accurate even though ANNs can assess the effect of all inputs simultaneously as well as SVMs can.

The inputs retained by the 15l feature selection are all used to build the SVMs, because they are the result of an accurate and proper feature selection procedure. Therefore, we can make some reflections on the environmental, physical-chemical and climate variables affecting *A. pallipes* presence in a more confident way.

> SELECTED VARIABLES

All the chemical-physical variables used for building the best-performing 15l models are variables that had already been reported to be important for *A. pallipes* distribution (Barbaresi *et al.*, 2007; Trouilhé *et al.*, 2007; Favaro *et al.*, 2010). In particular, the present data underlines the fact that the organic matter dissolved in the water – corresponding to the BOD₅ index – is an important factor for explaining white-clawed crayfish distribution in Piedmont (Favaro *et al.*, 2010). Moreover, water conductivity, the percentage of dissolved oxygen in water and the concentrations of PO₄³⁻, NO₃⁻ and NH₄⁺ are variables that were selected since they are indicative of water-body pollution. The importance of these variables for the species' presence in Piedmont has been discussed previously (Favaro *et al.*, 2010).

There are the variables that relate to the mineral component of the underlined geology. Among these, Ca²⁺ was selected since it is a key factor for determining the occurrence of crayfish. One reason is that a minimal concentration of 2.56 mg·L⁻¹ is essential for species exoskeleton calcification (Smith *et al.*, 1996). Another reason is that Ca²⁺ permits researchers to distinguish the limestone areas – those less suitable for crayfish – from the granite or the siliceous areas (Favaro *et al.*, 2010). As a consequence, pH was selected in that it relates to the calcium concentrations positively, as the water from areas with granite or siliceous rocks is less basic than the water from areas with limestone.

The environmental inputs that can explain species presence consist of the amount of bedrock and the amount of shade (due to canopy cover). Bedrock is important because it constitutes a stable habitat, mostly during periods of high flow. Shade is important because it provides relief during the hottest periods. The source of shade, riparian vegetation, provides shelter for crayfish – roots as well as branches and leaves that fall into streams. The number of shelters and burrows in a stream is a critical factor for the survival of adults – in fact, the most important resource bottleneck in crayfish populations (Hobbs, 1975, Barbaresi *et al.*, 2007).

The altitude and the minimum temperature during the coldest periods are essential for the presence of the species because these features are strictly connected with the thermal conditions of the sites. The precipitation during the wettest periods and water velocity are crucial for the success of *A. pallipes* populations (Nardi *et al.*, 2004) because they reflect the water flow regimen.

In conclusion, our approach underlines the synergistic effects of the many factors tied in with *A. pallipes* occurrence. Hence, it provides information vital for maintaining natural populations of indigenous crayfish and for selecting potential sites and streams where reintroduction strategies may be planned.

Moreover, the present research project suggests that SVMs are most likely a tool that is even more promising for freshwater management than traditional and other machine-learning techniques. However, much remains to be learned about the potential of SVMs for freshwater species conservation and many aspects of this issue deserve further investigation. We trust that the present study can provide a solid basis for future research.

ACKNOWLEDGEMENTS

Our research project was funded by the CRT (Cassa di Risparmio di Torino) foundation through the Alfieri project. We thank Vincent Marsicano for revising the English version.

REFERENCES

- Adriaenssens V., De Baets B., Goethals P.L.M. and De Pauw N., 2004a. Fuzzy rule-based models for decision support in ecosystem management. *Sci. Total Environ.*, 319, 1–12.
- Adriaenssens V., Goethals P.L.M., Charles J. and De Pauw N., 2004b. Application of Bayesian belief networks for the prediction of macroinvertebrate taxa in rivers. *Ann. Limnol. - Int. J. Lim.*, 40, 181–191.
- Baillie J. and Groombridge B., 1996. IUCN Red List of Threatened Animals, IUCN, Gland, 105 p.
- Barbaresi S., Cannicci S., Vannini M. and Fratini S., 2007. Environmental correlates of two macro-decapods distribution in Central Italy: Multi-dimensional ecological knowledge as a tool for conservation of endangered species. *Biol. Conserv.*, 36, 431–441.
- Broquet T., Thibault M. and Neveu A., 2002. Distribution and habitat requirements of the white clawed-crayfish, *Austropotamobius pallipes*, in a stream from the pays de Loire region, France: an experimental and descriptive study. *Bull. Fr. Pêche. Piscic.*, 367, 717–728.
- Brown M.P.S., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M. and Haussler D., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97, 262–267.
- Brusconi S., Bertocchi S., Renai B., Scalici M., Souty-Grosset C. and Gherardi F., 2008. Conserving indigenous crayfish: stock assessment and habitat requirements in the threatened *Austropotamobius italicus*. *Aquat. Cons. Mar. Freshw. Ecosyst.*, 18, 1227–1239.
- Burges C.J.C., 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 3, 121–167.
- Chan K.C. and Wong A.K., 1991. A statistical technique for extracting classificatory knowledge from databases. In: Piatetsky-Shapiro R. and Frawley W. (eds.), Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, 107–123.
- Cristianini N. and Scholkopf B., 2002. Support vector machines and kernel methods – the new generation of learning machines. *Ai Mag.*, 23, 31–41.
- Cohen J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20, 37–46.
- Cushing J.B. and Wilson T., 2005. Eco-informatics for decision makers advancing a research agenda. In: Ludäscher B. and Raschid L. (eds.), Data Integration in the Life Sciences: Second International Workshop, DILS 2005, San Diego, CA, USA, Proceedings, Lecture Notes in Computer Science, 3615, Springer-Verlag, Berlin, 325–334.
- D'heygere T., Goethals P.L.M. and De Pauw N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. *Ecol. Model.*, 195, 20–29.
- Dakou E., D'heygere T., Dedecker A.P., Goethals P.L.M., Lazaridou-Dimitriadou M. and De Pauw N., 2007. Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquat. Ecol.*, 41, 399–411.
- De'ath G. and Fabricius K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81, 3178–3192.
- Decoste D. and Scholkopf B., 2002. Training invariant support vector machines. *Mach. Learn.*, 46, 161–190.
- Dedecker A.P., Goethals P.L.M., Gabriëls 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. *Scientific World J.*, 2, 96–104.
- Dedecker A., Goethals P.L.M., Gabriëls W. and De Pauw N., 2004. Optimisation of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium). *Ecol. Model.*, 174, 161–173.
- Dedecker A.P., Goethals P.L.M. and 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. and Park Y.S. (eds.), Modelling Community Structure in Freshwater Ecosystems, Springer-Verlag, Berlin, 133–146.

- Dedecker A., Van Melckebeke K., Goethals P.L.M. and De Pauw N., 2007. Development of migration models for macroinvertebrates in the Zwalm river basin (Flanders, Belgium) as tools for restoration management. *Ecol. Model.*, 203, 72–86.
- Drake J.M., Randin C. and Guisan A., 2006. Modelling ecological niches with support vector machines. *J. Appl. Ecol.*, 43, 424–432.
- Duda R.O., Hart P.E. and Stork D.G., 2001. Pattern Classification, John Wiley & Sons, New York, 654 p.
- Dzeroski S., Demsar D. and Grbovic J., 2000. Predicting chemical parameters of river water quality from bioindicator data. *Appl. Intell.*, 13, 7–17.
- Elith J., Graham C.H., Anderson R.P., Dudik M., Ferrier S., Guisan A., Hijmans R.J., Huettmann F., Leathwick J.R., Lehmann A., Li J., Lohmann L.G., Loiselle B.A., Manion G., Moritz C., Nakamura M., Nakazawa Y., Overton J.McC., Peterson A.T., Phillips S.J., Richardson K.S., Scachetti-Pereira R., Schapire R.E., Soberón J., Williams S., Wisz M.S. and Zimmermann N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129–151.
- Favaro L., Tirelli T. and Pessani D., 2010. The role of water chemistry in the distribution of *Austropotamobius pallipes* (Crustacea Decapoda Astacidae) in Piedmont (Italy). *C. R. Biol.*, 333, 68–75.
- Ferrier S. and Guisan A., 2006. Spatial modelling of biodiversity at the community level. *J. Appl. Ecol.*, 43, 393–404.
- Fielding A.H., 1999. Machine Learning Methods for Ecological Applications, Kluwer Academic Publishers, New York, 280 p.
- Fielding A.H. and Bell J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.*, 24, 38–49.
- Foody G.M., 1992. On the compensation for chance agreement in image classification accuracy assessment. *Photogramm. Eng. Rem. S.*, 58, 1459–1460.
- Gabriels W., Goethals P.L.M., Dedecker A.P., Lek S. and De Pauw N., 2007. Analysis of macrobenthic communities in Flanders, Belgium, using a stepwise input variable selection procedure with artificial neural networks. *Aquat. Ecol.*, 41, 427–441.
- Gelder S.R., Delmastro G.B. and Rayburn J.N., 1999. Distribution of native and exotic branchiobdellidans (Annelida: Clitellata) on their respective crayfish hosts in northern Italy, with first record of native Branchiobdella species on an exotic North American crayfish. *J. Limnol.*, 58, 20–24.
- Gherardi F., Acquistapace P. and Santini G., 2001. Foraging in the white-clawed crayfish, *Austropotamobius pallipes* a threatened species. *Arch. Hydrobiol.*, 152, 339–351.
- Goethals P.L.M., Džeroski S., Vanrolleghem P. and De Pauw N., 2001. Prediction of benthic macroinvertebrate taxa (Asellidae and Tubificidae) in watercourses of Flanders by means of classification trees, IWA 2nd World water congress, Berlin, 5–6.
- Goethals P.L.M., Dedecker A.P., Gabriels W., Lek S. and De Pauw N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquat. Ecol.*, 41, 491–508.
- Goldberg D.E., 1989. Genetic Algorithm in Search, Optimization and Machine Learning, Addison-Winsley Publishing Company, Reading, 412 p.
- Grandjean F., Cornuault B., Archambault S., Bramard M. and Otrebsky G., 2000. Life history and population biology of the white-clawed crayfish, *Austropotamobius pallipes pallipes*, in a brook from the Poitou-Charentes region (France). *Bull. Fr. Pêche. Piscic.*, 356, 55–70.
- Green J.L., Hastings A., Arzberger P., Ayala F.J., Cottingham K.L., Cuddington K., Davis F., Dunne J.A., Fortin M.J., Gerber L. and Neubert M., 2005. Complexity in ecology and conservation: mathematical, statistical, and computational challenges. *BioScience*, 55, 501–510.
- Guisan A. and Zimmermann N.E., 2000. Predictive habitat distribution models in ecology. *Ecol. Model.*, 135, 147–168.
- Guo Q., Kellya M., and Graham C.H., 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecol. Model.*, 182, 75–90.
- Hall M.A., 1998. Correlation-based Feature Subset Selection for Machine Learning, Ph.D. Dissertation, University of Waikato, Waikato, New Zealand.
- Hoang H., Recknagel F., Marshall J. and Choy J., 2001. Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia). *Ecol. Model.*, 146, 195–206.
- Hoang H., Lock K., Mouton A. and Goethals P.L.M., 2010. Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecol. Inform.*, 5, 140–146.

- Hobbs Jr. H.H., 1975. Adaptations and convergence in North American crayfish. *Freshwater Crayfish*, 2, 541–551.
- Holdich D.M., 2003. Ecology of the white-clawed crayfish *Austropotamobius pallipes*, Conserving Natura 2000 Rivers Ecology Series No. 1, English Nature, Peterborough, 17 p.
- Holdich D.M. and Rogers W.D., 1997. Strategy for the management of white-clawed crayfish (*Austropotamobius pallipes*) populations in England and Wales, R&D Project 640, Environment Agency, Bristol, 23 p.
- Hosmer D. and Lemeshow S., 2000. Applied Logistic Regression, John Wiley and Sons Inc., New York, 392 p.
- Hu Q. and Davis C., 2005. Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine. *Mar. Ecol. Prog. Ser.*, 295, 21–31.
- Huang C., Davis L.S. and Townshend J.R.G., 2002. An assessment of support vector machines for land cover classification. *Int. J. Remote Sens.*, 23, 725–749.
- Joachims T., 1998. Text categorization with support vector machines: learning with many relevant features. *In: Proceedings of ECML-98, 10th European Conference on Machine Learning*, Springer-Verlag, Berlin, 137–142.
- Keerthi S.S., Shevade S.K., Bhattacharya C. and Murthy K.R.K., 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.*, 13, 637–649.
- Landis J.R. and Koch G.G., 1977. The measurements of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Lek S., Belaud A., Baran P., Dimopoulos I. and Delacoste M., 1996. Role of some environmental variables in trout abundance models using neural networks. *Aquat. Living Resour.*, 9, 23–29.
- Lencioni V., Maiolini B., Marziali L., Lek S. and Rossaro B., 2007. Macroinvertebrate assemblages in glacial stream systems: a comparison of linear multivariate methods with artificial neural networks. *Ecol. Model.*, 203, 119–131.
- Manel S., Dias J.M., Buckton S.T. and Ormerod S.J., 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *J. Appl. Ecol.*, 36, 734–747.
- Manel S., Williams H.C. and Ormerod S.J., 2001. Evaluating presence/absence models in ecology: the need to account for prevalence. *J. Appl. Ecol.*, 38, 921–931.
- Mouton A.M., De Baets B. and Goethals P.L.M., 2009. Knowledge-based versus data-driven fuzzy habitat suitability models for river management. *Environ. Model. Softw.*, 24, 982–993.
- Nardi P.A., Bernini F., Bo T., Bonardi A., Fea G., Ferrari S., Ghia D., Negri A., Razzetti E. and Rossi S., 2004. Il gambero di fiume nella provincia di Alessandria, PI-ME, Pavia, 111 p.
- Nardi P.A., Bernini F., Brocca M., Fea G., Ghia D. and Spairani M., 2006. Esperienze di introduzione di *Austropotamobius italicus* Ler. in un SIC del Parco regionale lombardo della valle del Ticino. *Pianura*, 20, 127–145.
- Nix H.A., 1986. A biogeographic analysis of Australian Elapid snakes. *In: Longmore R. (ed.), Atlas of Australian Elapid Snakes*, Australian Flora and Fauna Series, 8, 4–15.
- Nyström P., 1999. Ecological impact of introduced and native crayfish on freshwater communities: European perspectives. *In: Gherardi F. and Holdich D.M. (eds.), Crayfish in Europe as alien species: How to make the best of a bad situation?*, AA Balkema, Rotterdam, 63–85.
- Nyström P., Brönmark C. and Granéli W., 1999. Influence of an exotic and a native crayfish species on a littoral benthic community. *Oikos*, 85, 545–553.
- Obach M., Wagner R., Werner H. and Schmidt H.H., 2001. Modelling population dynamics of aquatic insects with artificial neural networks. *Ecol. Model.*, 146, 207–217.
- Olden J.D. and Jackson D.A., 2002. A comparison of statistical approaches for modelling fish species distributions. *Freshw. Biol.*, 47, 1976–1995.
- Park Y.S. and Chon T.S., 2007. Biologically-inspired machine learning implemented to ecological informatics. *Ecol. Model.*, 203, 1–7.
- Peterson A.T. and Vieglais D.A., 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *BioScience*, 51, 363–371.
- Pivard S., Demšar D., Lecomte J., Debeljak M. and Džeroski S., 2008. Characterizing the presence of oilseed rape feral populations on field margins using machine learning. *Ecol. Model.*, 212, 147–154.

- Platt J.C., 1998. Fast training of support vector machines using sequential minimal optimization. *In: Schölkopf B., Burges C. and Smola A. (eds.), Advances in Kernel Methods: Support Vector Machines*, MIT Press, Cambridge, 185–208.
- Platt J.C., 1999. Using sparseness and analytic QP to speed training of support vector machines. *In: Kearns M.S., Solla S.A. and Cohn D.A. (eds.), Advances in neural information processing systems*, 11, MIT Press, Cambridge, 557–563.
- Quinlan J.R., 1990. Decision trees and decision making. *IEEE Transactions on Systems, Man and Cybernetics*, 20, 339–346.
- Recknagel F., 2001. Application of machine learning to ecological modelling. *Ecol. Model.*, 146, 303–310.
- Recknagel F., 2003. *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation*, Springer-Verlag, Berlin and New York, 425 p.
- Ribeiro R. and Torgo L., 2008. A comparative study on predicting algae blooms in Douro River, Portugal. *Ecol. Model.*, 212, 86–91.
- Salski A. and Sperlbaum C., 1991. A fuzzy logic approach to modeling in ecosystem research. *In: Bouchon-Meunier B., Yager R.R., and Zadeh L.A. (eds.), Uncertainty in Knowledge Bases, 3rd International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU '90, Paris, France, July 2–6, 1990, Lecture Notes in Computer Science*, 521, Springer-Verlag, Berlin, 520–527.
- Sanchez-Hernandez C., Boyd D.S. and Foody G.M., 2007a. Mapping specific habitats from remotely sensed imagery: support vector machine and support vector data description based classification of coastal saltmarsh habitats. *Ecol. Inform.*, 2, 83–88.
- Sanchez-Hernandez C., Boyd D.S. and Foody G.M., 2007b. One-class classification for mapping a specific land-cover class: SVDD classification of Fenland. *IEEE Trans. Geosci. Remote Sens.*, 45, 1061–1073.
- Schleiter I.M., Obach M., Borchardt D. and Werner H., 2001. Bioindication of chemical and hydro-morphological habitat characteristics with benthic macro-invertebrates based on artificial neural networks. *Aquat. Ecol.*, 35, 147–158.
- Shan Y., Paull D. and McKay R.I., 2006. Machine learning of poorly predictable ecological data. *Ecol. Model.*, 195, 129–138.
- Smith G.R.T., Learner M.A., Slater F.M. and Foster J., 1996. Habitat features important for the conservation of the native crayfish *Austropotamobius pallipes* in Britain. *Biol. Conserv.*, 75, 239–246.
- Stockwell D.R.B. and Noble I.R., 1992. Induction of sets of rules from animal distribution data: a robust and informative method of analysis. *Math. Comput. Simul.*, 33, 385–390.
- Tax D.M.J. and Duijn R.P.W., 2004. Support vector data description. *Mach. Learn.*, 54, 45–66.
- Tirelli T. and Pessani D., 2009. Use of decision tree and artificial neural network approaches to model presence/absence of *Telestes muticellus* in Piedmont (North-Western Italy). *River Res. Appl.*, 24, 1001–1012.
- Tirelli T., Mussat Sartor R., Bona F., De Biaggi E., Zocco D. and Badino G., 2008. Census of *Austropotamobius* genus in four Districts of Piedmont (Western Italy). *Boll. Mus. Reg. Sci. Nat. Torino*, 25, 159–171.
- Tirelli T., Pozzi L. and Pessani D., 2009. Use of different approaches to model presence/absence of *Salmo marmoratus* in Piedmont (Northwestern Italy). *Ecol. Inform.*, 4, 234–242.
- Tirelli T., Favaro L. and Pessani D., submitted. Performance comparison among multivariate and data mining approaches to model presence/absence of *Austropotamobius pallipes* complex in Piedmont (Northwestern Italy).
- Trouilhé M.C., Souty-Grosset C., Grandjean F. and Parinet B., 2007. Physical and chemical water requirements of the white-clawed crayfish (*Austropotamobius pallipes*) in western France. *Aquat. Conserv.*, 17, 520–538.
- Vapnik V.N., 1995. *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 314 p.
- Witten I.H. and Frank E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn., Morgan Kaufmann Publishers, San Francisco, 371 p.