A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers

Knowledge and Management of Aquatic Ecosystems (2013) 409, 07 http://www.kmae-journal.org

DOI: 10.1051/kmae/2013052

A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers

E.J. Olaya-Marín[(1),*], F. Martínez-Capel[(1)], P. Vezza[(1)]

Received December 27, 2012
Revised May 5, 2013
Accepted May 7, 213

ABSTRACT

Key-words: *Artificial neural networks, random forests, native fish, species richness, Mediterranean rivers*

Machine learning (ML) techniques have become important to support decision making in management and conservation of freshwater aquatic ecosystems. Given the large number of ML techniques and to improve the understanding of ML utility in ecology, it is necessary to perform comparative studies of these techniques as a preparatory analysis for future model applications. The objectives of this study were (i) to compare the reliability and ecological relevance of two predictive models for fish richness, based on the techniques of artificial neural networks (ANN) and random forests (RF) and (ii) to evaluate the conformity in terms of selected important variables between the two modelling approaches. The effectiveness of the models were evaluated using three performance metrics: the determination coefficient (R^2), the mean squared error (MSE) and the adjusted determination coefficient (R^2_{adj}) and both models were developed using a k-fold crossvalidation procedure. According to the results, both techniques had similar validation performance (R^2 = 68% for RF and R^2 = 66% for ANN). Although the two methods selected different subsets of input variables, both models demonstrated high ecological relevance for the conservation of native fish in the Mediterranean region. Moreover, this work shows how the use of different modelling methods can assist the critical analysis of predictions at a catchment scale.

RÉSUMÉ

Une comparaison des réseaux de neurones et des forêts aléatoires pour prédire la richesse en espèces de poissons indigènes dans les rivières méditerranéennes

Mots-clés : *Réseaux de neurones, forêts aléatoires, poissons indigènes, richesse spécifique, rivières méditerranéennes*

Les techniques d'apprentissage automatique (ML) sont devenues importantes pour aider à la décision dans la gestion et la conservation des écosystèmes aquatiques d'eau douce. Étant donné le grand nombre de techniques ML pour améliorer la compréhension de l'utilité des ML en écologie, il est nécessaire de réaliser des études comparatives de ces techniques comme analyse préparatoire pour des applications de modèles futurs. Les objectifs de cette étude étaient : (i) de comparer la fiabilité et la pertinence écologique de deux modèles prédictifs pour la richesse de poisson, basé sur les techniques de réseaux de neurones artificiels (ANN) et les forêts aléatoires (RF) et (ii) d'évaluer la conformité en termes de sélection des variables importantes entre les deux approches de modélisation. L'efficacité des modèles a été évaluée au moyen de trois indicateurs de

(1) Institut d'Investigació per a la Gestió Integrada de Zones Costaneres, Universitat Politècnica de València, C/ Paranimf, 1, 46730 Grau de Gandia, València, Spain
* Corresponding author: estherjuliaolaya@gmail.com

Article published by EDP Sciences

performance : le coefficient de détermination (R^2), l'erreur quadratique moyenne (MSE) et le coefficient de détermination ajusté (R^2_{adj}) et les deux modèles ont été développés en utilisant une procédure de validation croisée k-fold. Selon les résultats, les deux techniques ont des performances de validation similaires (R^2 = 68 % pour RF et R^2 = 66 % pour ANN). Bien que les deux méthodes aient choisi différents sous-ensembles de variables d'entrée, les deux modèles ont démontré la pertinence écologique pour la conservation des poissons indigènes dans la région méditerranéenne. En outre, ce travail montre comment l'utilisation de différentes méthodes de modélisation peut aider à l'analyse critique des prévisions à l'échelle du bassin versant.

INTRODUCTION

In the last decades, due to the worldwide accelerated degradation of freshwater ecosystems (Beechie *et al.*, 2010; Strayer and Dudgeon, 2010) ecological modelling has become an important tool for wildlife and habitat conservation (Drew *et al.*, 2011). Particularly in Mediterranean rivers, pollution, introduction of exotic species and alteration of hydrological regimes have influenced fish population decline and, in some cases, the extinction of native species (García-Berthou *et al.*, 2005; Smith and Darwall, 2006). According to IUCN, 56% of freshwater Mediterranean species are threatened (Smith and Darwall, 2006) and, given the high degree of endemicity of biota and its high vulnerability to habitat alteration, more research is currently needed on local and native fish populations (Corbacho and Sánchez, 2001; Doadrio, 2002).

The conservation of fish diversity is one of the most critical issues facing the preservation of Mediterranean biodiversity (Smith and Darwall, 2006); and, due to its sensitivity to human disturbances, fish species richness is widely used as a primary indicator of ecological change and as a criterion for the selection of conservation areas (van Jaarsveld *et al.*, 1998; Lek *et al.*, 2005; He *et al.*, 2010). Increasing knowledge about the relationships between environmental features and fish populations is therefore essential for the design of effective habitat conservation and river restoration actions.

Ecological and biological data rarely satisfy the principles of parametric approaches, in which data must be independent, normal and homoscedastic (Guisan and Zimmermann, 2000; Breiman, 2001b) These criteria increase the challenges in modelling ecological phenomena. To cope with these issues, machine learning (ML) techniques have been widely used due to their ability to identify non-linear relationships and generate less uncertain predictive results (Olden *et al.*, 2008).

Several researchers have applied ML in ecological studies (Aertsen *et al.*, 2010; Armitage and Ober, 2010; Leclere *et al.*, 2011; Mouton *et al.*, 2011). In particular, artificial neural networks (ANN) and random forests (RF) are two machine learning techniques which are currently valuable tools for ecological modelling, and are especially useful in analysing large datasets and identifying non-linear relationships (Drew *et al.*, 2011). ANN are recognized as powerful and effective tools (Mastrorillo *et al.*, 1998; Olden *et al.*, 2008) to solve complex dependencies which are difficult with other traditional statistical methods (Lek *et al.*, 2005; Olden *et al.*, 2008). In the context of freshwater fish studies, ANN have been used with satisfactory results (Tirelli *et al.*, 2009; Olaya-Marín *et al.*, 2012). Ibarra *et al.* (2003) used ANN and multiple regression models (MLR) to identify the factors that influence fish guilds in the Garonne river basin (south-western France). They found better predictions of fish guilds with ANN than MLR. A similar result about ANN prediction accuracy is reported in Tirelli and Pessani (2009, 2011), who used ANN and decision trees to predict the presence of *Telestes muticellus* and *Alburnus alburnus alborella* in Piedmont rivers (north-western Italy). Moreover, Tirelli *et al.* (2009) applied ANN, discriminant function analysis, logistic regression and decision tree to model *Salmo marmoratus* distribution in Piedmont (Italy) and the performance of ANN was superior to the other modelling techniques. Also for imbalanced data, Hauser-Davis *et al.* (2010) concluded that ANN are an excellent alternative in classification problems.

Regarding RF, it is currently considered a promising technique in ecology (Cutler *et al.*, 2007; Franklin, 2010; Drew *et al.*, 2011; Cheng *et al.*, 2012) but it has rarely been applied in freshwater fish studies. RF can be used to identify important associations among variables (Evans *et al.*, 2011) and perform both regression and classification analyses (Cheng *et al.*, 2012). He *et al.* (2010) compared the use of classification and regression trees (CART) and RF to predict endemic fish assemblages and species richness in the upper Yangtze River. The study showed that RF is better than CART in terms of accuracy and efficiency. Knudby *et al.* (2010) used linear (GLM) and generalized additive models (GAM), Bagging, RF, Boosted Regression Trees (BRT) and support vector machines (SVM) to build predictive mapping of reef fish species richness, diversity and biomass. They found that the tree-based models were generally superior to predict species richness of reef fish. Furthermore, Mouton *et al.* (2011) found similar predictive performance of RF and Fuzzy logic models to represent mesohabitat suitability for *Salmo trutta* in Spain, whereas Kampichler *et al.* (2010) compared different ML techniques (including ANN and RF) for classification problems and recommend the use of RF in conservation biology.

Given the large number of ML techniques, there are not established guidelines for defining the most appropriate method to address a particular ecological question or management action for freshwater ecosystems. In particular, ML regression models are very scarce in ecology (Franklin, 2010) resulting in the necessity of conducting comparative studies of ML techniques. Moreover, the ecological knowledge of Mediterranean rivers needs to be expanded. Additional efforts to improve the understanding of the main factors influencing species richness are valuable (Filipe *et al.*, 2010; Aparicio *et al.*, 2011). In this context, the objectives of this study were (i) to compare the reliability and ecological relevance of two ML predictive models for fish richness, based on the techniques of ANN and RF and (ii) to evaluate their conformity in terms of selected important variables between the two modelling approaches. It is important to highlight here that a comparison between ANN and RF for prediction of fish species richness has not yet been presented in the literature. These comparisons are currently considered a new open line of research (Aertsen *et al.*, 2011) and this paper represents a further contribution in such a field.

MATERIALS AND METHODS

> *STUDY AREA AND DATA COLLECTION*

This study was carried out with data collected in the mainstems of the Júcar, Cabriel and Turia Rivers, in the Eastern Iberian Peninsula (Figure 1). These rivers are characterized by a Mediterranean climate, a flow regime controlled by rainfall variability, a strong seasonal and inter-annual discharge variation with two high flow periods per year (spring and fall) and severe droughts in summer (Ollero *et al.*, 2011). The mean temperature ranges between 11.6 to 17 °C and the maximum temperatures are registered in July and August, coinciding with the dry period (Estrela *et al.*, 2004). The mean annual precipitation in the study area is 500 mm, ranging between 320 mm in dry years to 800 mm in the wet years (Estrela *et al.*, 2004). The soils are highly permeable and are characterized by high infiltration and percolation rates (Estrela *et al.*, 2004). During the last decades, the natural flow regime has been altered by the construction of reservoirs and water diversion structures; flow regulation is severe particularly for streams located in the middle and lower part of the watersheds. The effect of flow regulation is expressed by an inversion of the intra-annual variability pattern; in summer, the regulated flow is greater than natural flow, and in contrast, the regulated flow is smaller than the natural flow during winter (Aparicio *et al.*, 2011). Due to industrial and urban waste water, pollution also affects rivers (Estrela *et al.*, 2004; Aparicio *et al.*, 2011) and agricultural practices, particularly in spring and summer, constitute a source of diffusive pollutants at the catchment-scale (Estrela *et al.*, 2004).

In the analyses, we used data from 90 sampling sites along the mainstems of the three rivers (Figure 1). The sites were selected as representative in terms of river morphology

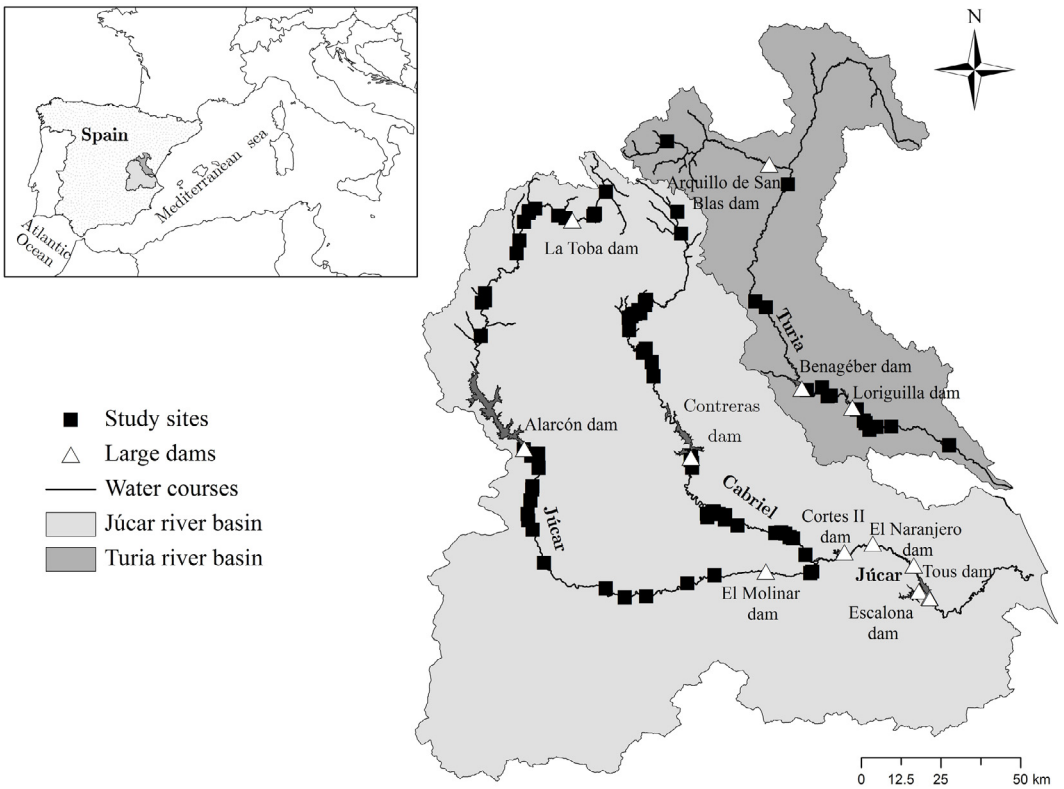

Figure 1
Study area showing the distribution of the 90 sampling sites in the three rivers studied (Júcar, Cabriel and Turia rivers).

Table I
Freshwater fish species present in the study area related to its threat status (Freyhof and Brooks, 2011; IUCN, 2012).CR, critically endangered; EN, endangered; VU, vulnerable; NT, near threatened; LC, least concern.

Species name	Common name	Family	Threat status
Anguilla anguilla	European eel	Anguillidae	CR
Parachondrostoma arrigonis	Júcar nase	Cyprinidae	CR
Parachondrostoma turiense	Turia nase	Cyprinidae	EN
Achondrostoma arcasii	Bermejuela	Cyprinidae	VU
Barbus haasi	Iberian redfin barbel	Cyprinidae	VU
Cobitis paludica	Southern Iberian spined-loach	Cobitidae	VU
Luciobarbus guiraonis	Eastern Iberian barbel	Cyprinidae	VU
Squalius pyrenaicus	Southern Iberian chub	Cyprinidae	NT
Squalius valentinus	Eastern Iberian chub	Cyprinidae	VU
Iberocypris alburnoides	Calandino	Cyprinidae	VU
Salmo trutta	Brown trout	Salmonidae	LC
Salaria fluviatilis	Freshwater blenny	Blenniidae	LC

and proportion of mesohabitats which characterize the analysed water courses. Native fish species richness (*i.e.* the number of fish species at each sampling site) was defined by means of a single-pass electrofishing during the spring/summer period from 2005 to 2009. The limits of the sampling sites were opened and the minimum length of each sampled reach was 50 m.

The total fish diversity comprises 12 native species (Table I) with a maximum local richness of five species. These values are common in Mediterranean rivers, which are generally characterized by a low species richness per site (Ferreira *et al.*, 2007). Cyprinidae is the predominant

Table II

Potential environmental variables used to build the predictive models for native fish species richness. Physico-chemical and hydrological parameters were obtained from the monitoring network (MN) of the Júcar river basin authority, mean width and hydro-morphological unit proportions were measured in situ during fish samplings, while geographical variables were derived from GIS analyses.

Variable	Code	Source	Mean	Standard deviation
Dissolved oxygen (mg· L^{-1})	DIS	MN	9.58	0.44
Biological oxygen demand (mg· L^{-1})	BOD	MN	2.51	0.77
Total phosphorus (mg· L^{-1})	TOP	MN	0.06	0.03
Nitrite (mg· L^{-1})	NIT	MN	0.02	0.02
pH	PH	MN	8.18	0.11
Suspended solids (mg· L^{-1})	SUS	MN	11.39	5.77
Water conductivity (μS· cm^{-1})	CON	MN	797.87	172.62
Water temperature (°C)	WAT	MN	13.38	2.48
Percentage of pools (%)	POO	*in situ*	48.66	21.42
Percentage of glides (%)	GLI	*in situ*	11.21	16.89
Percentage of riffles (%)	RIF	*in situ*	28.27	21.41
Percentage of rapids (%)	RAP	*in situ*	5.79	6.85
Percentage of runs (%)	RUN	*in situ*	6.07	12.10
Mean width of the water surface (m)	WID	*in situ*	12.46	4.68
Channel length without artificial barriers (km)	CWB	GIS	26.35	29.46
Altitude (m a.s.l)	ALT	GIS	746.48	298.43
Drainage area (km^2)	DRA	GIS	3318.84	2607.51
Distance from the source (km)	DHS	GIS	150.19	76.41
Mean annual flow rate (m^3· s^{-1})	FMA	MN	4.33	2.44
Inter-annual mean flow (calculated for 5 years) (m^3· s^{-1})	FIA	MN	5.50	2.57
Coefficient of variation of mean monthly flows (referred to fish sampling)	FIM	MN	0.58	0.18
Coefficient of variation of mean annual flows (calculated for 5 years)	FCV	MN	0.40	0.17
Index of riparian habitat quality	QBR	MN	73.61	20.74
Iberian biomonitoring working party	IBMWP	MN	131.68	36.32

family in the three rivers; the most important genera are *Achondrostoma, Parachondrostoma, Luciobarbus, Barbus, Squalius* and *Iberocypris.* Other species present in the rivers are *Cobitis paludica* and *Salaria fluviatilis* which are very sensitive to pollution and have distinct environmental requirements (CHJ, 2007). All these species perform small-scale migrations for reproduction within the river system and the only one migrating at a large scale is *Anguilla Anguilla,* a catadromous fish species with a complex life-history that includes migrations across the Atlantic Ocean. The number of individuals of these native fish species have decreased consistently in the last decades as a consequence of habitat modifications (including barriers) and pollution in the lower river reaches (Doadrio, 2001; Costa *et al.*, 2012).

24 environmental variables (Table II) were used to construct the ANN and RF models, which were selected considering their ecological importance for fish life cycle (Granado-Lorencio, 1996; Jackson *et al.*, 2001; Bernardo *et al.*, 2003; Costa *et al.*, 2012). These variables belong to three categories: physicochemical water quality, hydro-morphology, and biologically-based indicators of water and riparian habitat quality. Following Olden *et al.* (2006), we took into account predictive variables from multiple spatial scales (*i.e.* mesohabitat, river segment and catchment scale). This multiscale modelling approach allowed an integrative analysis of multiple sources of variability and a better understanding of the biodiversity patterns in Mediterranean rivers (Filipe *et al.*, 2010).

Data were collected from three main sources: *in situ*, GIS analyses and from the Júcar River Basin monitoring network (MN) (Table II). Physico-chemical variables (*i.e.* dissolved oxygen, biological oxygen demand, total phosphorous, nitrite, pH, suspended solids, water temperature) are consistent with the mean annual reported for the survey year. The proportions

of hydro-morphological units (HMUs) and the mean width of water surface were measured *in situ*. The classification of HMUs was based on the methods proposed by Dolloff *et al*. (1993) using five types of mesohabitats (pool, glide, riffle, rapid and run) (Alcaraz-Hernández *et al.*, 2011):

- Pools: water depth >0.6 m, water velocity below the average for the reach, and a very low longitudinal gradient.
- Glides: water depth <0.6 m, water velocity similar to the average for the reach, little turbulence and nearly symmetrical cross sections.
- Riffles: shallow water with ripples on the surface, an average water velocity < 0.4 m·s^{-1}, nearly symmetrical crosssections, and a mean depth similar to the mean substrate size.
- Rapids: shallow water with water velocity greater than the average for the reach, very high energy dissipation, elements of coarse substrate projecting from the water surface, steep channel gradients.
- Runs: moderate channel slopes with depths higher than riffle moderate to fast current, surface not turbulent, and button materials ranging from small gravel to rubble.

Geographical variables (*i.e.* channel length without artificial barriers, altitude, drainage area and distance from the source) were delineated using ArcGIS 9.3.1 software (ESRI©2009). The mean monthly flow was calculated at ungauged sites through a linear interpolation between gauged sites. To define the hydrological indexes (inter-annual mean flow and the coefficients of variation of mean monthly flow and of mean annual flow), we used the linear relationship between the natural flow and the accumulated drainage area, and then transformed monthly flow values to regulated conditions (Leopold *et al.*, 1964; Caissie, 2006). The riparian habitat quality index (QBR, Munné *et al.*, 2003) was taken into account to assess the morphological conditions of the sampling sites; this index was adopted by the Spanish Ministry of Environment (MMARM, 2008). QBR consists of four components, which synthesize qualitative features related to the conservation state of the riparian area: total vegetation cover, vegetation cover structure and quality, and river channel alterations. The values of this index are distributed in five quality intervals (⩾95: excellent quality; 90-75: good quality; 70-55: moderate quality; 503: poor quality; ⩽ 25: bad quality). Finally, we used the Iberian Biomonitoring Working Party index (IBMWP) (Alba-Tercedor, 1996) based on invertebrate analysis to evaluate the biological quality of rivers. IBMWP values are distributed in five ranges of water quality: ⩾101: very clean water; 100-61: unpolluted water or not appreciably altered; 60-36: partially polluted water with some evident effects; 35-16: very polluted water; ⩽15: heavily polluted water.

> MODELLING TECHNIQUES

Artificial neural networks (ANN)

Artificial neural networks are mathematical models inspired by the structure and behaviour of the human brain (Olden *et al.*, 2008). They are considered a powerful computational tool to address ecological issues that are difficult to analyse by traditional statistical methods (Lek *et al.*, 2005); among different types of ANN, multilayer perceptron (MLP) is the most used in ecology (Özesmi *et al.*, 2006). It is constituted by multiple layers and the information is transferred from input layer to the output (feed-forward). This kind of ANN is based on supervised learning, which implies the use of input and output datasets to iteratively change the weights until the simulated outputs are similar to the observed ones. To minimize the error, the algorithm employs the values of the error calculated in the previous iteration and then updates the weights. A detailed description of ANN is reported in Olden *et al.* (2008) and Goethals *et al.* (2007).

This study applied a MLP to predict native fish species richness. We built and tested several MLP models to establish, by trial and error estimates, the optimal number of neurons in the hidden layer. A single hidden layer was used to significantly reduce the computational time.

Moreover, as reported in Kurková (1992) the use of a single hidden layer produces similar results compared to the incorporation of additional hidden layers. Before training, data were scaled proportionally between –1 and 1 in the range of the minimum and maximum values (Demuth *et al.*, 2010).

The transfer function in the hidden layer was a hyperbolic tangent and an identity function in the case of the output layer. The hyperbolic tangent gave optimal results in previous studies (Isa *et al.*, 2010) in which a performance comparison was carried out to select the best MLP activation function (Olaya-Marín *et al.*, 2012). The Levenberg-Marquardt (LM) optimization algorithm was used to train the candidate models because this algorithm is the fastest method to train neural networks of moderate size (Karul *et al.*, 2000). LM has been applied successfully in ecology (Gutiérrez-Estrada and Bilton, 2010) and a description of the algorithm can be found in Singh *et al.* (2009). To test and validate the models, we used the *k*-fold cross-validation procedure and tested different *k* values (ranging from 3 to 10, Goethals *et al.*, 2007; Dormann, 2011). The best *k* value was then identified by the comparison of the performance of the different ANN obtained in the cross-validation procedure. This approach is frequently used when the number of data available is not sufficient to divide the data into training and validation datasets (Goethals *et al.*, 2007; Olden *et al.*, 2008) The dataset is randomly split in *k* subsets: *k*-1 folds are used for training and the remaining used for validation (Hastie *et al.*, 2009) This procedure is repeated *k* times and results in *k* models. The performance of the different *k* models was averaged as final criterion for model evaluation. All numerical ANN calculations were performed using MATLAB software (version R2010a).

Random forests

The fish richness was also predicted using the random forests (RF) methodology (Breiman, 2001a; Cutler *et al.*, 2007) in the statistical software R (R Development Core Team, 2009) by means of the randomForest package (Liaw and Wiener, 2002). RF is an ensemble learning technique based on a combination of a large set of decision trees. Each tree is trained by selecting random bootstrap samples X_i (i = bootstrap iteration which ranges from 1 to t) of the original dataset X and a random set of predictive variables. This process represents the main difference compared to standard decision trees (Breiman *et al.*, 1984), where each node is split using the best split among all predictive variables (*e.g.* Vezza *et al.*, 2012). Moreover, RF corrects many of the known issues in CART, such as over-fitting (Breiman, 2001a; Cutler *et al.*, 2007), and provides very well-supported predictions with large numbers of independent variables (Cutler *et al.*, 2007).

As the response variable (fish richness) was numerical, we confined our attention to regression RF models. The algorithm for growing a RF of t regression trees was performed as follows (for full details see Breiman, 2001a):

(1) t bootstrap samples X_i of the training dataset were randomly drawn with replacement, each one containing approximately two-thirds of the elements of the original dataset X. The elements not included in each training dataset are referred to as out-of-bag data (OOB, *i.e.* the validation dataset) for that bootstrap sample. On average, each element of X was an OOB element in one-third of the t iterations.

(2) For each bootstrap sample X_i, an unpruned regression tree was grown. At each node m variables were randomly selected and the best split was automatically chosen.

(3) New data (OOB elements) were predicted by averaging the predictions of the generated t trees. In particular, for each element (y_i) of the original dataset an aggregated prediction (g_{OOB}) was developed and the out-of-bag estimate of the error rate (E_{OOB}) was computed as:

$$\left[E_{OOB} = (1/t)\sum_{1}^{t}[y_i - g_{OOB}(X_i)]^2\right] \cdot$$

The E_{OOB} was also used to choose an optimal value of t and m (Breiman, 2001a). The m parameter (number of variables permutated at each node) was defined as [1/3 × (number of

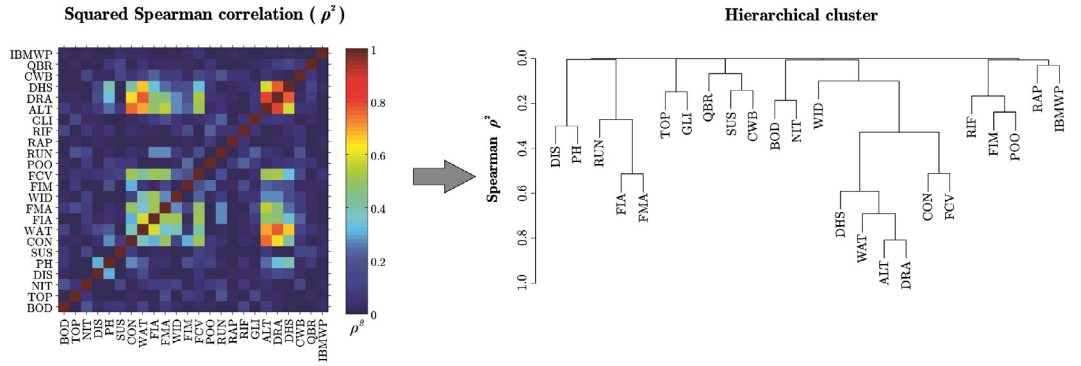

Figure 2
Correlation matrix and hierarchical clustering using squared Spearman correlation (ρ^2) of potential predictors.

variables)] with a minimum of $m = 2$ (see Breiman, 2001a). As E_{OOB} is an unbiased estimate of the generalization error, it is not necessary to test the predictive ability of the model using a cross-validation procedure (Breiman, 2001a). However, in accordance with ANN and for a more reliable comparison, we performed *k*-fold cross-validation (with *k* ranging from 3 to 10) following the approach reported in Hastie *et al*. (2009).

Variable selection

As a first step, a correlation matrix was calculated to verify collinearity (Figure 2). For high correlations (Spearman's rho^2 $>$ 0.5) we removed the variable (Table II) with the least ecological relevance (Dormann, 2011) This choice was based on the authors' expert knowledge and on the findings of previous research studies (Aparicio *et al.*, 2000; Vila-Gispert *et al.*, 2005; Alcaraz-Hernández *et al.*, 2011; Hermoso and Clavero, 2011). To identify the most important predictive variables we followed two different approaches. On one hand, the forward stepwise procedure was applied in the ANN models; this method consists of adding step by step a single input variable and then evaluating the improvement in ANN performance (Gevrey *et al.*, 2003). The irrelevant input variables are therefore eliminated measuring the complexity reduction of the ANN model (see Gevrey *et al.*, 2003; Tirelli and Pessani, 2009) and at the end of the process the variables that imply a significant improvement in the ANN performances are selected.

On the other hand, we applied the model improvement ratio technique (MIR, Murphy *et al.*, 2010) to identify the most parsimonious RF model. RF produces a measure of variable importance by analysing the deterioration of the predictive ability of the model when each predictor is replaced in turn by random noise. The increase in the mean squared error of each tree (IncMSE) is used as a score of importance of a given variable (Vincenzi *et al.*, 2011), as it indicates the contribution to RF prediction accuracy for that variable. The MIR technique uses the variable importance standardized from zero to one and the improvement ratio was therefore calculated as [*In*/*Imax*], where *In* is the importance of a given variable and *Imax* is the maximum model improvement score. We then iterated through MIR thresholds (*i.e.* 0.05 increments), with all variables above the threshold retained for each model (Evans and Cushman, 2009). The models corresponding to different subsets were then compared and the model exhibiting the minimum MSE error and the lowest number of variables was selected.

Model evaluation

Model comparison was conducted under two different conditions: (i) before variable selection to show models' efficiencies and variable importance using all sources of variability and their

interactions; and (ii) after variable selection as described in Section "Variable selection", to evaluate the effect of variables selection in RF and ANN.

The overall accuracy of the two statistical models was evaluated using three performance metrics, commonly used in ecological modelling (*e.g.* Singh *et al.*, 2009; Aertsen *et al.*, 2011): the determination coefficient (R^2), the mean squared error (MSE) and the adjusted determination coefficient (R^2_{adj}).

The determination coefficient (R^2) assesses the proportion of variability explained by the model, and it is calculated by:

$$R^2 = \left(\frac{\sum \left(Y^{sim} \cdot Y^{obs}\right) - \left(\left(\sum Y^{sim} \cdot \sum Y^{obs}\right)/n\right)}{\sqrt{\left(\sum Y^{sim^2} - \left(\left(\sum Y^{sim}\right)^2/n\right)\right) \cdot \left(\sum Y^{obs^2} - \left(\left(\sum Y^{obs}\right)^2/n\right)\right)}} \right)^2 \quad (1)$$

where, Y^{obs} are the observed values, Y^{sim} represent the simulated values, and n is the total number of observations.

MSE is the error between model predictions and observed values, it is expressed as:

$$\text{MSE} = \frac{1}{n} \sum \left(Y^{sim} - Y^{obs}\right)^2 \quad (2)$$

The adjusted determination coefficient is a modification of the determination coefficient and was used during the model selection procedures to compare models with different numbers of predictive variables (Vezza *et al.*, 2010) In contrast to R^2, this coefficient penalizes the excessive use of inputs, and it is expressed as follows:

$$R^2_{adj} = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1} \quad (3)$$

where p represents the number of input variables.

Finally, the ecological interpretation of each optimal ANN and RF model was carried out by sensitivity analysis and the assessment of the relative importance of the inputs. The partial derivatives (PaD) method was applied in ANN (Dimopoulos *et al.*, 1995), which represents the mostly used approach to evaluate the relative importance in MLP (Gevrey *et al.*, 2003) The PaD method can be used to estimate the sensitivity of the output as a function of small variations in each input variable; positive PaD values indicate a positive relationship between the corresponding input variable and the output variable, and *vice versa*. Also PaD is useful estimating input relative importance. On the other hand, partial plots were calculated to visualize the marginal effect of predictive variables in RF simulations and its relative importance was indicated by the IncMSE values (Breiman, 2001a).

RESULTS

> *ENVIRONMENTAL VARIABLE*

The variables DHS, WAT, ALT, and DRA had a high Spearman's rank correlation coefficient (Figure 2; see Table II for variable codes). According to Filipe *et al.* (2010) and Oberdorff *et al.* (1995) DRA is an important variable to estimate the distribution of species richness. For this reason, we removed DHS, WAT and ALT as potential predictors. CON and FCV were also highly correlated; we kept FCV as a potential predictive variable because it represents flow variability and it was found important for Mediterranean fish species in several studies (Magalhães *et al.*, 2007; Hermoso and Clavero, 2011) Lastly, FIA and FMA were also correlated; however both of them were used as predictive variables given their similar relevance for Mediterranean fish life cycle (Hermoso and Clavero, 2011).

Table III
Relative importance of input variables in ANN and RF before variable selection. See Table II for variable codes.

Ranking	ANN model (20-15-1)		RF model	
	Variable	Relative importance (%)	Variable	Relative importance (%)
1	IBMWP	14.80	QBR	10.41
2	RIF	14.79	RUN	10.05
3	FIM	8.63	DRA	9.56
4	CWB	6.95	IBMWP	8.89
5	RAP	6.58	RAP	7.78
6	DRA	6.45	SUS	6.66
7	QBR	5.46	RIF	5.87
8	FMA	5.26	PH	5.76
9	GLI	4.51	CWB	5.46
10	DIS	3.92	DIS	4.89
11	FCV	3.20	FMA	4.07
12	PH	3.02	FIM	3.85
13	SUS	3.00	GLI	3.40
14	BOD	2.67	POO	3.32
15	RUN	2.63	FCV	2.94
16	POO	2.09	WID	2.43
17	NIT	1.72	FIA	2.21
18	FIA	1.70	BOD	1.83
19	WID	1.48	NIT	0.42
20	TOP	1.15	TOP	0.21

> MODEL RESULTS BEFORE VARIABLE SELECTION

ANN model showed an efficiency of 0.19 MSE in training and 0.65 in validation using 20 environmental variables as predictors. On the other hand, RF had a training MSE of 0.1 and a validation MSE of 0.68 The first six most important variables to simulate fish species richness in ANN model were IBMWP, RIF, FIM, CWB, RAP and DRA. In contrast, the most significant variables in the RF model were QBR, RUN, DRA, IBMWP, RAP and SUS (Table III).

The best neural network architecture to predict native fish richness had three layers (*i.e.* 7-6-1), with seven neurons in the input layer (which corresponds to the predictive variables), a hidden layer with six neurons, and the output layer with a single neuron; the last one calculates the estimated values of native fish species richness. During the *k*-fold crossvalidation, the ANN performance did not increase with *k* values higher than 6, therefore we used $k = 6$ to validate the model. The stepwise selection of variables and the MSE is illustrated in Figure 3. In RF, the OOB error stabilization occurred between $t = 1500$ and $t = 2500$ replicates. However, a heuristic estimation of *t* taking into account the OOB error stabilization was defined as [2·(*t* for E_{OOB} stabilization) = 5000] (Evans and Cushman, 2009). The relation between the MSE and number of variables in RF is shown in Figure 3. For both models, the MSE quickly decreased as the number of input variables was increasing (Figure 3). A breakpoint was located at 7 variables in ANN; and at 5 variables in RF. Based on this criterion, we used 7 predictive input variables to build the ANN model and 5 for RF.

> MODEL RESULTS AFTER VARIABLE SELECTION

According to the correlation analysis and the forward stepwise procedure the relevant variables to predict the native fish richness with the ANN model are reported, in order of importance: IBMWP, RIF, FMA, FIM, CWB, DRA and QBR. In contrast for RF, the selected variables were QBR, RUN, DRA, IBMWP and RAP.

The best training performance was obtained by the RF model ($R^2 = 0.94$; MSE = 0.10), whereas for validation both techniques gave similar results (MSE = 0.62, $R^2 = 66\%$ for ANN;

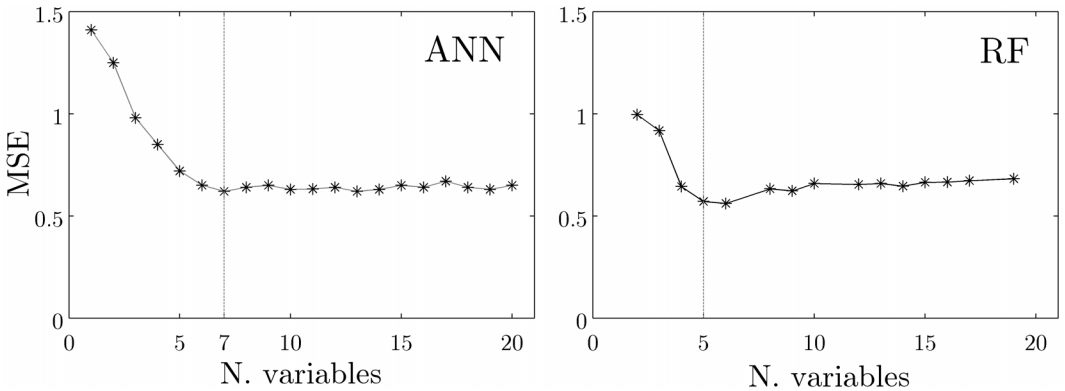

Figure 3
Artificial neural networks (ANN) and random forests (RF) performance in terms of mean squared error (MSE) as a function of the number of input variables (N. variables). The final ANN model (including 7 variables) and RF model (including 5 variables) were those in which the incorporation of any additional variable meant no relevant error decrease (vertical lines).

Table IV
Predictive performances for the training and validation set of the 6-fold cross-validation for native fish species richness. Models were evaluated with determination coefficient (R^2), mean squared error (MSE) and adjusted determination coefficient (R^2_{adj}).

Models		Training			Validation		
		R^2	MSE	R^2_{adj}	R^2	MSE	R^2_{adj}
ANN	Min	0.76	0.30	0.74	0.52	0.44	0.48
	Max	0.84	0.43	0.82	0.77	0.87	0.75
	Mean	**0.81**	**0.35**	**0.78**	**0.66**	**0.62**	**0.63**
	SD	0.03	0.05	0.03	0.08	0.19	0.09
RF	Min	0.93	0.09	0.93	0.58	0.38	0.57
	Max	0.95	0.12	0.95	0.84	0.74	0.78
	Mean	**0.94**	**0.10**	**0.94**	**0.68**	**0.56**	**0.66**
	SD	0.01	0.01	0.01	0.11	0.25	0.15

MSE = 0.56, R^2 = 68% for RF). Table IV displays the performance of training and validation in RF and ANN models, while Figure 4 shows the scatter plots between observed and simulated values.

The implementation of the partial derivatives algorithm for ANN revealed that the most important variables to predict native fish richness were IBMWP, with a relative importance of 20.72% and percentage of riffle (RIF) with an importance of 20.18%. In the case of the RF model, the most important variables were QBR and percentage of runs (RUN), with a relative importance of 23.51% and 22.02%, respectively (Figure 5). Both models have in common three variables (IBMWP, QBR and DRA), following the sensitivity analysis we found a positive relationship between IBMWP and QBR with native fish richness in both models (*i.e.* an increase in these variables leads to increments of richness in the study area). In contrast, the richness is negatively related to DRA in both RF and ANN models (Figure 5).

DISCUSSION

In this study two machine learning techniques (*i.e.* ANN and RF) were applied to estimate the fish species richness in the Júcar river basin, as preparatory reference for future wildlife and habitat conservation actions. The methodology compared the reliability and ecological relevance of the two statistical techniques in order to evaluate their applicability and assess the conformity in terms of variables importance between the two predictive models.

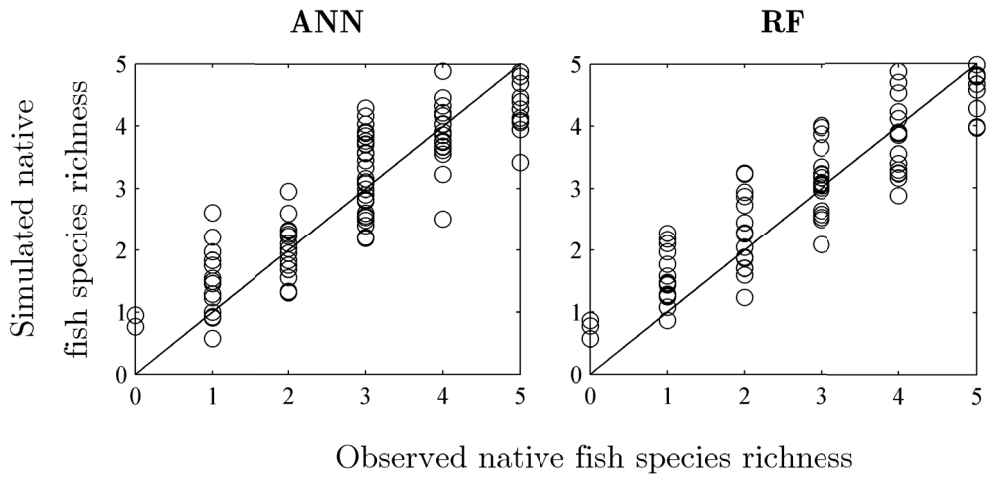

Figure 4
Scatter plots between the observed and the simulated values of native fish species richness for the ANN and RF models.

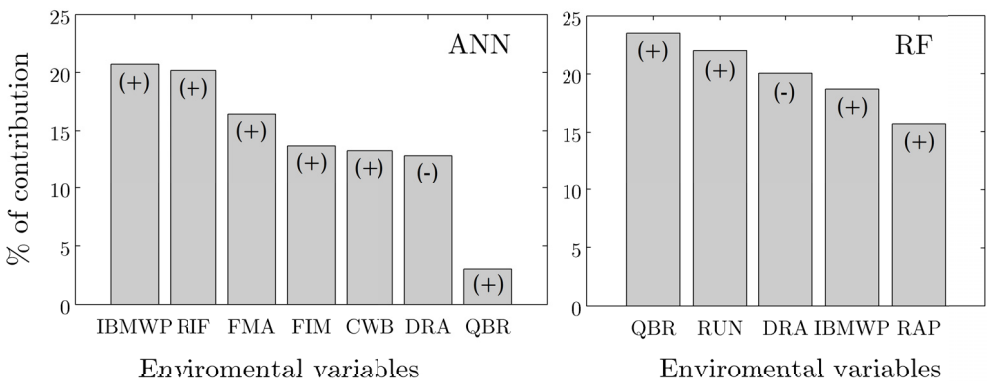

Figure 5
Relative importance (expressed in % of contribution) of each input variable to predict native fish richness. Left side: the ANN model, right side: the RF model. See Table II for variable codes.

The ANN model built before variable selection had a high difference between training and validation performances due to the presence of large sources of variability which make the learning process difficult and could generate problems of over-fitting (Maier and Dandy, 2000). However, the most important variables in this ANN model (*e.g.* IBMWP, RIF, FIM and CWB, see Table III) can be considered ecologically relevant to fish. Similarly, the RF model built with 20 variables also shows a notable difference in training and validation performances. Also for the RF case, the most important predictors in this model have been identified by previous studies as relevant to freshwater fish species (Bernardo *et al.*, 2003; Costa *et al.*, 2012). The high differences between training and validation performances confirm the hypothesis that a very complex model with many degrees of freedom is not robust (Maier and Dandy, 2000)

Looking at the models built after variable selection, ANN and RF showed no significant differences of performance in the cross-validation procedure (R^2 = 68% for RF and R^2 = 66% for ANN, Table IV). However, it is important to note that RF outperformed ANN in terms of MSE, particularly considering small numbers of input variables (Figure 3, *N*. Variables < 7).

The RF model had the smallest number of inputs and only five variables were required for prediction, while ANN required seven. The difference in the number of inputs highlighted the advantage of using RF, because the models with fewer variables are much easier to interpret and can reduce the level of prediction uncertainty (Jorgensen and Fath, 2011). However, the

RF model showed much higher accuracy in the training compared to that obtained in the validation phase (Table IV), presenting a considerable difference in performance. In contrast, for ANN the difference between training and validation prediction error was smaller and demonstrated more stable results.

Since the 90's, diverse mathematical algorithms have emerged in order to quantify and interpret the importance and contribution of input variables to the model output and, at same time, to identify and eliminate redundant variables to increase model parsimony (*e.g.* Olden and Jackson, 2002; Gevrey *et al.*, 2003; Murphy *et al.*, 2010). In this research we used PaD (Dimopoulos *et al.*, 1995) for ANN and the model improvement ratio (Murphy *et al.*, 2010) for RF. PaD allowed the classification of the input variables according to their contribution to the output variable and, in accordance with Gevrey *et al.* (2003), the technique produced stable variables ranking over the different ANN models. On the other hand, MIR demonstrated to be a simple and powerful methodology to select the threshold that minimized both retained inputs and model error.

For variable selection, a forward stepwise methodology was embedded in the ANN algorithm, whereas the variable importance values were used in RF to screen the overall range of inputs and select the most parsimonious model. The two procedures were based on different approaches and led to two different sets of variables. However, this result is not surprising and is confirmed in several studies (Xu and Zhang, 2001; Abrahamsson *et al.*, 2003; Reunanen, 2003; Wells *et al.*, 2011), in which different variable selection procedures produced similar subsets of variables. It is important to note that the RF ranking of variables is based on all possible combinations of model inputs with *m* random variables permuted at each node of the tree. In contrast, the one-step-ahead search procedure of ANN may not lead to the best combination of inputs; it required the modeller to study the sequence of variables and analyse whether the addition or removal of a few more variables might not produce any improvement.

Another important aspect in ecological modelling involves the evaluation and interpretation of the results. The presented models were in accordance with the literature due to the fact that the selected input variables, such as water quality, flow regime and the status of riparian forest, are of great importance for the Mediterranean fish populations (Granado-Lorencio, 1996, 2000; Bernardo *et al.*, 2003; Ferreira *et al.*, 2007).

Both models had three variables in common to predict fish richness: drainage area (DRA), quality index of the riparian forest (QBR) and the biological index for water quality (IBMWP). Although the variables' ranking was not the same (in terms of % of contribution, Figure 5), this accordance can indicate the robustness of the results (*e.g.* Xu and Zhang, 2001) In several studies (Oberdorff *et al.*, 1995; Reyjol *et al.*, 2007; Leprieur *et al.*, 2009) DRA is considered an important environmental variable for fish species richness, integrating information related to habitat diversity, the variety of microclimates and flow regimes in the river basin (Allan and Castillo, 2007; Townsend *et al.*, 2008) The negative relationship between DRA and richness may be due to the fact that the lower reaches of Júcar, Cabriel and Turia rivers are highly regulated and have higher levels of contamination than the upper reaches (Aparicio *et al.*, 2011).

Water quality and riparian forest play a key role in determining the richness of native fish Thus, QBR and IBMWP have been identified as relevant factors positively influencing fish species richness in Spain (Carballo *et al.*, 2009; Sánchez-Montoya *et al.*, 2010) and are widely used for the ecological monitoring of rivers. Indeed, the riparian forest provide shelter and food for aquatic organisms (Naiman *et al.*, 1993) and can strongly influence the quality of aquatic habitats, particularly along a gradient of river regulation (Garófano-Gómez *et al.*, 2011). Furthermore, river pollution is currently one of the most important threats for the Mediterranean freshwater fish (Smith and Darwall, 2006); it can severely disrupt the functioning of the aquatic ecosystem and compromise the survival of biota (Granado-Lorencio, 2000).

Both ANN and RF selected the proportion of HMUs as important predictors of fish richness and, in particular, percentage of riffles (RIF) were selected in ANN and percentage of runs (RUN) and rapids (RAP) in RF, all showing a positive relationship with the target variable. Although the two statistical techniques considered different HMU types, one can observe

how the spatial distribution and dynamics of mesohabitats can be of great importance for fish conservation (Fausch *et al.*, 2002). According to Bernardo *et al.* (2003) riffles and runs can influence the composition of Mediterranean fish communities; particularly for those dominated by the family Cyprinidae (Granado-Lorencio, 2000; Ferreira *et al.*, 2007), because these mesohabitats can offer good conditions in terms of food availability and shelters.

Mean annual flow rate (FMA), the coefficient of variation of mean monthly flow (FIM) along with the channel length without artificial barriers (CWB) were only selected by ANN. Different studies highlighted the role of flow variability and magnitude in supporting the aquatic communities (Poff *et al.*, 1997; Belmar *et al.*, 2011; Olaya-Marín *et al.*, 2012) and, the longitudinal connectivity has important consequences on the distribution of native fish (Kroes *et al.*, 2006.), either small or large-scale migratory species (*e.g. Parachondrostoma arrigonis* or *Anguilla Anguilla,* both critically endangered in the region of interest). Including these aspects in fish richness prediction underline the ecological relevance of the ANN model, which seemed to capture the interplay between natural and anthropogenic factors influencing fish species distribution.

As reported in Siroky (2009) RF models are fast to train and tune. In our research the time needed for RF model construction was much less than for ANN (few minutes compared to hours) due to the structure of RF algorithm characterized by few parameters to set and a limited number of variables (*m*) to be permutated at each tree node. ANN needed more time for computer architecture design and learning as it performs a large amount of trials changing the number of neurons and the type of activation function in the hidden layer. However, the amount of time needed can be reduced by using different computer processors working in parallel (Armitage and Ober, 2010) and once calibrated, ANN are able to process a large volume of data and quickly generate predictions (Olden *et al.*, 2008).

The applied ML techniques involve elegant mathematical theories and are known to be robust to noise and able to manage the non-linearity among variables (Lek *et al.*, 2005; Olden *et al.*, 2008; Siroky, 2009), but for some authors they can be seen as black boxes (Hooten, 2011). In particular, the relationships between the input variables and the predicted values produced by ANN and RF do not have simple representations such as a formula (*e.g.* linear regressions) or pictorial graph (*e.g.* regression trees) that characterizes the entire function, and this lack of simple representation can make the ecological application difficult (Cutler *et al.*, 2007). Olden and Jackson (2002) provided an interesting point of view to give light into the so called "black box"; but, compared to traditional statistical methods, ML remain more complex to understand and apply (Olden and Jackson, 2002; Olden *et al.*, 2008). In addition, these techniques require the modeller to have knowledge and experience in designing and programming the algorithms in order to make effective use of the tools and to reach satisfactory and valid results (Franklin, 2010) In spite of the emergence of high level programming languages and user-friendly toolboxes the modeller must have knowledge and expertise in building these kind of statistical models (Aertsen *et al.*, 2010; Franklin, 2010).

Although a comprehensive evaluation of several different techniques was beyond the scope of this paper, we believe that the best predictive ML method cannot be chosen *a priori* and both ANN and RF constitute valuable tools to predict fish richness in the Mediterranean region. Looking at the results, we can state that the use of more than one ML technique on the same study area was helpful, not only to identify the most important variables, but also to interpret the goodness and coherence of the results. As an operational procedure for future studies on fish species richness, we can state that the comparison of different ML methods should be carried out. Moreover, as a further step already planned for the near future, these analyses can be performed in other Mediterranean basins.

The presented approaches, which relate environmental variables to the fish communities, can be used for predicting fish richness at the basin scale and can be incorporated into the decision-making process for water resources management (Paredes-Arquiola *et al.*, in press). For instance they could contribute to perform large-scale assessments of environmental flow standards, based on methodological frameworks with a regional perspective (Poff *et al.*, 2010; Paredes-Arquiola *et al.*, 2013).

ACKNOWLEDGEMENTS

This study was partially funded by the Spanish Ministry of Economy and Competitiveness with the projects SCARCE (Consolider-Ingenio 2010 CSD2009-00065) and POTECOL "Evaluación del Potencial Ecológico de Ríos Regulados por Embalses y Desarrollo de Criterios para su mejora según la Directiva Marco del Agua" (CGL2007-66412). In addition, the RF analysis was developed in the frame of the EU-funded HolRiverMed project (IEF, Marie Curie Actions). We thank the Confederación Hidrográfica del Júcar (Spanish Ministry of Agriculture, Food and Environment) for the data provided to develop this study and we also owe our gratitude to Sasa Plestenjak for the collaboration in building the first fish database for this research. We owe our gratitude to Chris Holmquist-Johnson and Leanne Hanson (USGS, Fort Collins Science Center) for the scientific review of the paper.

REFERENCES

Abrahamsson C., Johansson J., Sparén A. and Lindgren F., 2003. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets. *Chemometrics Intell. Lab. Syst.*, *69*, 3–12.

Aertsen W., Kint V., van Orshoven J., Özkan K. and Muys B., 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecol. Model.*, *221*, 1119–1130.

Aertsen W., Kint V., Van Orshoven J. and Muys B., 2011. Evaluation of modelling techniques for forest site productivity prediction in contrasting ecoregions using stochastic multicriteria acceptability analysis (SMAA). *Environ. Modell. Softw.*, *26*, 929–937.

Alba-Tercedor A., 1996. Macroinvertebrados acuaticos y calidad de las aguas de los ríos, IV Simposio del Agua en Andalucía (SIAGA), Almería, 203–213.

Alcaraz-Hernández J.D., Martínez-Capel F., Peredo-Parada M. and Hernández-Mascarell A.B., 2011. Mesohabitat heterogeneity in four mediterranean streams of the Jucar river basin (Eastern Spain). *Limnetica*, *30*, 363–378.

Allan J.D. and Castillo M.M., 2007. Stream ecology: structure and function of running waters, 2nd edn., Springer, Netherlands, 436 p.

Aparicio E., Vargas M.J., Olmo J.M. and de Sostoa A., 2000. Decline of native freshwater fishes in a Mediterranean watershed on the Iberian Peninsula: A quantitative assessment. *Environ. Biol. Fishes*, *59*, 11–19.

Aparicio E., Carmona-Catot G., Moyle P.B. and García-Berthou E., 2011. Development and evaluation of a fish-based index to assess biological integrity of Mediterranean streams. *Aquat. Conserv.: Mar. Freshwat. Ecosyst.*, *21*, 324–337.

Armitage D.W. and Ober H.K., 2010. A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecol. Inform.*, *5*, 465–473.

Beechie T.J., Sear D.A., Olden J.D., Pess G.R., Buffington J.M., Moir H., Roni P. and Pollock M.M., 2010. Process-based principles for restoring river ecosystems. *Bioscience*, *60*, 209–222.

Belmar O., Velasco J. and Martinez-Capel F., 2011. Hydrological classification of natural flow regimes to support environmental flow assessments in Intensively regulated Mediterranean Rivers, Segura River Basin (Spain). *Environ. Manage.*, *47*, 992–1004.

Bernardo J.M., Ilhéu M., Matono P. and Costa A.M., 2003. Interannual variation of fish assemblage structure in a Mediterranean river: implications of streamflow on the dominance of native or exotic species. *River Res. Appl.*, *19*, 521–532.

Breiman L., 2001a. Random Forests. *Mach. Learn.*, *45*, 5–32.

Breiman L., 2001b. Statistical modeling: the two cultures. *Stat. Sci.*, *16*, 199–231.

Breiman L., Friedman J., Olshen R. and Stone C., 1984. Classification and Regression Trees, Wadsworth International Group, Belmont, California, 368 p.

Caissie D., 2006. River discharge and channel width relationships for New Brunswick rivers. Canadian Technical Report of Fisheries and Aquatic Sciences, Rept. 2637, 26 p.

Carballo R., Cancela J., Iglesias G., Marín A., Neira X. and Cuesta T., 2009. WFD indicators and definition of the ecological status of rivers. *Water Resour. Manag.*, *23*, 2231–2247.

Cheng L., Lek S., Lek-Ang S. and Li Z., 2012. Predicting fish assemblages and diversity in shallow lakes in the Yangtze River basin. *Limnologica*, *42*, 127–136.

CHJ, 2007. Estudio general sobre la Demarcación Hidrográfica del Júcar, Confederación Hidrográfica del Júcar, Madrid, 206 p.

Corbacho C. and Sánchez J.M., 2001. Patterns of species richness and introduced species in native freshwater fish faunas of a Mediterranean-type basin: the Guadiana River (southwest Iberian Peninsula). *Regul. River.*, *17*, 699–707.

Costa R.M.S., Martínez-Capel F., Muñoz-Mas R., Alcaraz-Hernández J.D. and Garófano-Gómez V., 2012. Habitat suitability modelling at mesohabitat scale and effects of dam operation on the endangered Júcar nase, *Parachondrostoma arrigonis* (river Cabriel, Spain). *River Res. Appl.*, *28*, 740–752.

Cutler D.R., Edwards T.C., Beard K.H., Cutler A., Hess K.T., Gibson J. and Lawler J.J., 2007. Random Forests for classification in ecology. *Ecology*, *88*, 2783–2792.

Demuth H., Beale M. and Hagan M., 2010. Neural network toolbox user's guide, The MathWorks Inc, Natick, Massachusetts, 901 p.

Dimopoulos Y., Bourret P. and Lek S., 1995. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Process. Lett.*, *2*, 1–4.

Doadrio I., 2001. Atlas y libro rojo de los peces continentales de España, Ministerio de Medio Ambiente, Madrid, 358 p.

Doadrio I., 2002. Origen y Evolución de la Ictiofauna Continental Española. *En*: Atlas y libro rojo de los peces continentales de España. 2da ed, CSIC y Ministerio del Medio Ambiente, Madrid, 20–34.

Dolloff C.A., Hankin D.G. and Reeves G.H., 1993. Basinwide estimation of habitat and fish populations in streams, U.S. Department of Agriculture, Blacksburg, Virginia, 25 p.

Dormann C.F., 2011. Modelling species' distributions. *In*: Jopp F., Reuter H. and Breckling B. (eds.), Modelling complex ecological dynamics: an Introduction into ecological modelling for students, teachers and scientists, Springer-Verlag, Berlin, 179–196.

Drew C.A., Wiersma Y. and Huettmann F., 2011. Predictive species and habitat modeling in landscape ecology: concepts and applications, Springer, New York, 328 p.

Estrela T., Fidalgo A., Fullana J., Maestu J., Pérez M.A. and Pujante A.M., 2004. Júcar Pilot River Basin, provisional article 5 report Pursuant to the Water Framework Directive, Confederación Hidrográfica del Júcar, Valencia, 200 p.

Evans J. and Cushman S., 2009. Gradient modeling of conifer species using random forests. *Landsc. Ecol.*, *24*, 673–683.

Evans J.S., Murphy M.A., Holden Z.A. and Cushman S.A., 2011. Modeling species distribution and change using Random Forest. *In*: Drew C.A., Wiersma Y.F. and Huettmann F. (eds.), Predictive Species and Habitat Modeling in Landscape Ecology, Springer New York, 139–159.

Fausch K., Torgersen C., Baxter C. and Li H., 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *Bioscience*, *52*, 483–498.

Ferreira T., Oliveira J., Caiola N., De Sostoa A., Casals F., Cortes R., Economou A., Zogaris S., Garcia de Jalón D., Ilhéu M., Martinez-Capel F., Pont D., Rogers C. and Prenda J., 2007. Ecological traits of fish assemblages from Mediterranean Europe and their responses to human disturbance. *Fisheries Manag. Ecol.*, *14*, 473–481.

Filipe A.F., Magalhães M.F. and Collares-Pereira M.J., 2010. Native and introduced fish species richness in Mediterranean streams: the role of multiple landscape influences. *Divers. Distrib.*, *16*, 773–785.

Franklin J., 2010. Mapping species distributions: spatial inference and prediction, Cambridge University Press, New York, 338 p.

García-Berthou E., Alcaraz C., Pou-Rovira Q., Zamora L., Coenders G. and Feo C., 2005. Introduction pathways and establishment rates of invasive aquatic species in Europe. *Can. J. Fish. Aquat. Sci.*, *62*, 453–463.

Garófano-Gómez V., Martínez-Capel F., Peredo-Parada M., Olaya-Marín E.J., Muñoz-Mas R., Costa R. and Pinar- Arenas L., 2011. Assessing hydromorphological and floristic patterns along a regulated Mediterranean river: The Serpis River (Spain). *Limnetica*, *30*, 307–238.

Gevrey M., Dimopoulos I. and Lek S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.*, *160*, 249–264.

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

Granado-Lorencio C., 1996. Ecología de peces, Universidad de Sevilla, Sevilla, 353 p.

Granado-Lorencio C., 2000. Ecología de comunidades: el paradigma de los peces de agua dulce, Universidad de Sevilla, Sevilla, 284 p.

Guisan A. and Zimmermann N.E., 2000. Predictive habitat distribution models in ecology. *Ecol. Model.*, *135*, 147–186.

Gutiérrez-Estrada J.C. and Bilton D.T., 2010. A heuristic approach to predicting water beetle diversity in temporary and fluctuating waters. *Ecol. Model.*, *221*, 1451–1462.

Hastie T., Tibshirani R. and Friedman J., 2009. The Elements of Statistical Learning: data mining, Inference and prediction, Springer, 768 p.

Hauser-Davis R.A., Oliveira T.F., Silveira A.M., Silva T.B. and Ziolli R.L., 2010. Case study: Comparing the use of nonlinear discriminating analysis and Artificial Neural Networks in the classification of three fish species: acaras (*Geophagus brasiliensis*), tilapias (*Tilapia rendalli*) and mullets (*Mugil liza*). *Ecol. Inform.*, *5*, 474–478.

He Y., Wang J., Lek-Ang S. and Lek S., 2010. Predicting assemblages and species richness of endemic fish in the upper Yangtze River. *Sci. Total Environ.*, *408*, 4211–4220.

Hermoso V. and Clavero M., 2011. Threatening processes and conservation management of endemic freshwater fish in the Mediterranean basin: a review. *Mar. Freshwater Res.*, *62*, 244–254.

Hooten M.B., 2011. The state of spatial and spatio-temporal statistical modeling. *In*: Drew C., Wiersma Y. and Huettmann F. (eds.), Predictive Species and Habitat Modeling in Landscape Ecology, Springer New York, 29–41.

Ibarra A.A., Gevrey M., Park Y.-S., Lim P. and Lek S., 2003. Modelling the factors that influence fish guilds composition using a back-propagation network: assessment of metrics for indices of biotic integrity. *Ecol. Model.*, *160*, 281–290.

Isa I.S., Omar S., Saad Z. and Osman M.K., 2010. Performance comparison of different multilayer perceptron network activation functions in automated weather classification. Proceedings of the 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, Kota Kinabalu, Malaysia, 71–75.

Jackson D.A., Peres-Neto P.R. and Olden J.D., 2001. What controls who is where in freshwater fish communities the roles of biotic, abiotic, and spatial factors. *Can. J. Fish. Aquat. Sci.*, *58*, 157–170.

Jorgensen S.E. and Fath B.D., 2011. Fundamentals of ecological modelling: applications in environmental management and research. 4th ed., Elsevier, Amsterdam, 432 p.

Kampichler C., Wieland R., Calmé S., Weissenberger H. and Arriaga-Weiss S., 2010. Classification in conservation biology: a comparison of five machine-learning methods. *Ecol. Inform.*, *5*, 441–450.

Karul C., Soyupak S., Çilesiz A.F., Akbay N. and Germen E., 2000. Case studies on the use of neural networks in eutrophication modeling. *Ecol. Model.*, *134*, 145–152.

Knudby A., LeDrew E. and Brenning A., 2010. Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sens. Environ.*, *114*, 1230–1241.

Kroes M.J., Gough P.P., Wanningen H., Schollema P., Ordeix M. and Vesely D., 2006. From sea to source. Practical guidance for the restoration of fish migration in European Rivers. Interreg IIIC Project “Community Rivers”, Groningen, The Netherlands, 119 p.

Kurková V., 1992. Kolmogorov’s theorem and multilayer neural networks. *Neural Netw.*, *5*, 501-506.

Leclere J., Oberdorff T., Belliard J. and Leprieur F., 2011. A comparison of modeling techniques to predict juvenile 0+ fish species occurrences in a large river system. *Ecol. Inform.*, *6*, 276–285.

Lek S., Scardi M., Verdonschot P., Descy J.P. and Park Y.S. (eds.), 2005. Modelling community structure in freshwater ecosystems, Springer-Verlag, Berlin.

Leopold L.B., Wolman M.G. and Miller J.P., 1964. Fluvial processes in geomorphology, W.H. Freeman, San Francisco, 544 p.

Leprieur F., Brosse S., García-Berthou E., Oberdorff T., Olden J.D. and Townsend C.R., 2009. Scientific uncertainty and the assessment of risks posed by non-native freshwater fishes. *Fish. Fish.*, *10*, 88–97.

Liaw A. and Wiener M., 2002. Classification and regression by Random Forest. *R News*, *2*, 18–22.

Magalhães M.F., Beja P., Schlosser I.J. and Collares-Pereira M.J., 2007. Effects of multi-year droughts on fish assemblages of seasonally drying Mediterranean streams. *Freshw. Biol.*, *52*, 1494–1510.

Maier H.R. and Dandy G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environ. Modell. Softw.*, *15*, 101–124.

Mastrorillo S., Dauba F., Oberdorff T., Guégan J.-F. and Lek S., 1998. Predicting local fish species richness in the garonne river basin. *C.R. Acad. Sci. - Ser. III - Sciences de la Vie*, *321*, 423–428.

MMARM, 2008. Orden MARM/2656/2008 de 10 septiembre, por la que se aprueba la instrucción de planificación hidrológica. BOE núm. 229, de 22 de septiembre de 2008., Ministerio de Medio Ambiente, y Medio Rural y Marino (MMARM), Madrid.

Mouton A.M., Alcaraz-Hernández J.D., De Baets B., Goethals P.L.M. and Martínez-Capel F., 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. *Environ. Modell. Softw.*, *26*, 615–622.

Munné A., Prat N., Solà C., Bonada N. and Rieradevall M., 2003. A simple field method for assessing the ecological quality of riparian habitat in rivers and streams: QBR index. *Aquat. Conserv.: Mar. Freshwat. Ecosyst.*, *13*, 147–163.

Murphy M.A., Evans J.S. and Storfer A., 2010. Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. *Ecology*, *91*, 252–261.

Naiman R.J., Decamps H. and Pollock M., 1993. The role of riparian corridors in maintaining regional biodiversity. *Ecol. Appl.*, *3*, 209–212.

Oberdorff T., Guégan J.-F. and Hugueny B., 1995. Global scale patterns of fish species richness in rivers. *Ecography*, *18*, 345–352.

Olaya-Marín E.J., Martínez-Capel F., Soares Costa R.M. and Alcaraz-Hernández J.D., 2012. Modelling native fish richness to evaluate the effects of hydromorphological changes and river restoration (Júcar River Basin, Spain). *Sci. Total Environ.*, *440*, 95–105.

Olden J.D. and Jackson D.A., 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.*, *154*, 135–150.

Olden J.D., Poff N.L. and Bledsoe B.P., 2006. Incorporating ecological knowledge into ecoinformatics: An example of modeling hierarchically structured aquatic communities with neural networks. *Ecol. Inform.*, *1*, 33–42.

Olden J.D., Lawler J.J. and Poff N.L., 2008. Machine learning methods without tears: A primer for ecologists. *Q. Rev. Biol.*, *83*, 171–193.

Ollero A., Ibisate A., Gonzalo L., Acín V., Ballarín D., Díaz E., Gimeno M., Domenech S., Granado D., García H., Mora D. and Sánchez M., 2011. The IHG index for hydromorphological quality assessment of rivers and streams: updated version *Limnetica*, *30*, 255–262.

Özesmi S.L., Tan C.O. and Özesmi U., 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecol. Model.*, *195*, 83–93.

Paredes-Arquiola J., Martinez-Capel F., Solera A. and Aguilella V., 2013. Implementing environmental flows in complex water resources systems–case study: the Duero river basin, Spain. *River Res. Appl.*, *29*, 451–468.

Paredes-Arquiola J., Solera-Solera A., Martínez-Capel F., Momblanch-Benavent A. and Andreu-Álvarez J. Integrating water management, habitat modelling and water quality at basin scale environmental flow assessment – Tormes River (Spain). *Hydrol. Sci. J.-J. Sci. Hydrol.*, in press.

Poff N.L., Allan J.D., Bain M.B., Karr J.R., Prestegaard K.L., Richter B.D., Sparks R.E. and Stromberg J.C., 1997. The natural klow regime. *Bioscience*, *47*, 769–784.

Poff N.L., Richter B.D., Arthington A.H., Bunn S.E., Naiman R.J., Kendy E., Acreman M., Apse C., Bledsoe B.P., Freeman M.C., Henriksen J., Jacobson R.B., Kennen J.G., Merritt D.M., O'Keeffe J.H., Olden J.D., Rogers K., Tharme R.E. and Warner A., 2010. The ecological limits of hydrologic alteration (ELOHA): a new framework for developing regional environmental flow standards. *Freshw. Biol.*, *55*, 147–170.

R Development Core Team, 2009. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 409 p.

Reunanen J., 2003. Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res.*, *3*, 1371–1382.

Reyjol Y., Hugueny B., Pont D., Bianco P.G., Beier U., Caiola N., Casals F., Cowx I., Economou A., Ferreira T., Haidvogl G., Noble R., De Sostoa A., Vigneron T. and Virbickas T., 2007. Patterns in species richness and endemism of European freshwater fish. *Glob. Ecol. Biogeogr.*, *16*, 65–75.

Sánchez-Montoya M.M., Vidal-Abarca M.R. and Suárez M.L., 2010. Comparing the sensitivity of diverse macroinvertebrate metrics to a multiple stressor gradient in Mediterranean streams and its influence on the assessment of ecological status. *Ecol. Indic.*, *10*, 896–904.

Singh K.P., Basant A., Malik A. and Jain G., 2009. Artificial neural network modeling of the river water quality–A case study. *Ecol. Model.*, *220*, 888–895.

Siroky D.S., 2009. Navigating Random Forests and related advances in algorithmic modeling. *Statist. Surv.*, *3*, 147–163.

Smith K.G. and Darwall W.R.T., 2006. The status and distribution of freshwater fish endemic to the mediterranean basin, IUCN – The World Conservation Union, Gland, Switzerland/Cambridge, UK., 41 p.

Strayer D.L. and Dudgeon D., 2010. Freshwater biodiversity conservation: recent progress and future challenges. *J. N. Am. Benthol. Soc.*, *29*, 344–358.

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.*, *25*, 1001–1012.

Tirelli T. and Pessani D., 2011. Importance of feature selection in decision-tree and artificial-neural-network ecological applications. *Alburnus alburnus alborella*: A practical example. *Ecol. Inform.*, *6*, 309-315.

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.

Townsend C., Begon M. and Harper J., 2008. Essentials of Ecology, 3rd edn, Wiley-Blackwell, Oxford.

van Jaarsveld A.S., Freitag S., Chown S.L., Muller C., Koch S., Hull H., Bellamy C., Kruger M., Endrody-Younga S., Mansell M.W. and Scholtz C.H., 1998. Biodiversity assessment and conservation strategies. *Science*, *279*, 2106–2108.

Vezza P., Comoglio C., Rosso M. and Viglione A., 2010. Low flows regionalization in North-Western Italy. *Water Resour. Manag.*, *24*, 4049–4074.

Vezza P., Parasiewicz P., Rosso M. and Comoglio C., 2012. Defining minimum environmental flows at regional scale: application of mesoscale habitat models and catchments classification. *River Res. Appl.*, *28*, 675–792.

Vila-Gispert A., Alcaraz C. and García-Berthou E., 2005. Life-history traits of invasive fish in small Mediterranean streams. *Biol. Invasions*, *7*, 107-116–116.

Vincenzi S., Zucchetta M., Franzoi P., Pellizzato M., Pranovi F., De Leo G.A. and Torricelli P., 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice lagoon, Italy. *Ecol. Model.*, *222*, 1471–1478.

Wells B., Yu C., Koroukian S. and Kattan M., 2011. Comparison of variable selection methods for the generation of parsimonious prediction models for use in clinical practice. In: Proceedings of the 33rd Annual Meeting of the Society for Medical Decision Making, Chicago, US.

Xu L. and Zhang W.-J., 2001. Comparison of different methods for variable selection. *Anal. Chim. Acta*, *446*, 475–481.