Can Multilayer Perceptron Ensembles model the ecological niche of freshwater fish species?

R. Muñoz-Mas[a*], F. Martínez-Capel[a], J.D. Alcaraz-Hernández[a], A.M. Mouton[b]

[a]Institut d'Investigació per a la Gestió Integrada de Zones Costaneres (IGIC), Universitat Politècnica de València, C/Paranimf 1, 46730 Grau de Gandia, València, Spain

[b]Research Institute for Nature and Forest (INBO), Kliniekstraat 25, B-1070 Brussels, Belgium

* Corresponding author. Tel.: (+34) 962849458; fax: (+34) 962849309

E-mail addresses: pitifleiter@hotmail.com (R. Muñoz-Mas), fmcapel@dihma.upv.es (F. Martínez-Capel), jdalcaraz@gmail.com (J.D. Alcaraz-Hernández), ans.mouton@inbo.be (A.M. Mouton)

Abstract

The potential of Multilayer Perceptron (MLP) Ensembles to explore the ecology of freshwater fish species was tested by applying the technique to redfin barbel (*Barbus haasi* Mertens, 1925), an endemic and montane species that inhabits the North-East quadrant of the Iberian Peninsula. Two different MLP Ensembles were developed. The physical habitat model considered only abiotic variables, whereas the biotic model also included the density of the accompanying fish species and several invertebrate predictors. The results showed that MLP Ensembles may outperform single MLPs. Moreover, active selection of MLP candidates to create an optimal subset of MLPs can further improve model performance. The physical habitat model confirmed the redfin barbel preference for middle-to-upper river segments whereas the importance of depth confirms that redfin barbel prefers pool-type habitats. Although the biotic model showed higher uncertainty, it suggested that redfin barbel, European eel and the considered cyprinid species have similar habitat requirements. Due to its high predictive performance and its ability to deal with model uncertainty, the MLP Ensemble is a promising tool for ecological modelling or habitat suitability prediction in environmental flow assessment.

Keywords

Artificial neural networks, *Barbus haasi,* data mining, species distribution modelling, uncertainty analysis.

1 Introduction

Ecological models for the quantitative prediction of species distributions are key to understanding the realised niche of species and its implication for species conservation in relation to global change (Austin, 2007). Therefore, ecological models have increasingly received attention due to their wide management applications in the context of biogeography, conservation biology and climate change studies (Mouton et al., 2010). Many studies on ecological modelling have focused on explanation rather than prediction (Elith and Leathwick, 2009); however, differences in the life-history or in the gene flow of fish assemblages could result in different realised niches (Mouton et al., 2010). Abiotic factors, together with dispersal and biotic interactions, are often considered the three elements that shape the ecological niche by determining species distribution and abundance (Barve et al., 2011). However, ecological models have usually focused on abiotic factors only (Boulangeat et al., 2012), and very few studies in freshwater fish ecology have explicitly included biotic variables (Elith and Leathwick, 2009) to explore biotic interactions and consumer-resource dynamics (Soberón, 2007). The consideration of these three elements (i.e. abiotic, biotic and dispersal factors) do not allow for simple statistical analysis because the data collected often exhibit non-linear and complex data structures (Crisci et al., 2012). Consequently, there is a need for new and innovative approaches to understand the complex structure of living systems (Larocque et al., 2011).

Several sophisticated modelling techniques have been applied in the ecological modelling of fish species, ranging from linear to multivariate and machine learning techniques such as Artificial Neural Networks (ANN) (Brosse and Lek, 2000, Muñoz-Mas et al., 2014, Palialexis et al., 2011). The most popular ANN architecture has been the Multilayer Perceptron (MLP) paradigm because it is considered to be able to approximate any continuous function (Olden et al., 2008). Formerly, MLP was referred to as a 'black box' because it provided little explanatory insight into the relative influence of variables in the prediction process (Olden and Jackson, 2002). To date, an enormous effort has been made to develop methods that clarify variable importance and interactions (Gevrey et al., 2006, Lek et al., 1996, Olden and Jackson, 2002), and consequently, MLPs should no longer be treated as 'black box' models (Özesmi et al., 2006).

There are several examples of single MLP applications in freshwater fish ecology (Park and Chon, 2007). For instance, MLPs have been successfully applied to model fish ecology through a broad range of ecosystems (Brosse and Lek, 2000, Gevrey et al., 2006, Kemp et al., 2007, Laffaille et al., 2003) and in some cases outperforming other statistical approaches (Baran et al., 1996, Lek et al., 1996). Despite those successful studies, it has been demonstrated that single models (e.g. a single MLP) do not necessarily perform consistently, resulting in divergent predictions (Buisson et al., 2010, Fukuda et al., 2011, Fukuda et al., 2013). The use of model ensembles has been emphasised to overcome this phenomenon (Araújo and New, 2007). The Multilayer Perceptron Ensemble (MLP Ensemble, Hansen and Salamon, 1990) has proven to be proficient in several areas of ecology (Palialexis et al., 2011, Watts and Worner, 2008), but has rarely been applied in freshwater ecosystems (Muñoz-Mas et al., 2014).

Fish communities in Mediterranean rivers are an interesting targets to develop these novel statistical approaches (Hopkins II and Burr, 2009), particularly communities dominated by cyprinids, as they are characterised by a high number of endemic species for which there is insufficient knowledge about their ecology (Ferreira et al., 2007). Furthermore, endemic species tend to facilitate a more robust analysis of species–environment relationships. In this paper, we focused on the redfin barbel (*Barbus haasi* Mertens, 1925), a rheophilic small barbel (maximum body-length 30 cm) that is endemic to the Iberian Peninsula (Bianco, 1998) and categorised as vulnerable (Freyhof and Brooks, 2011). Their populations have decreased markedly, with pollution and the presence of exotic species being the main factors involved in the decline (Perea et al., 2011). Although redfin barbel has been the subject of numerous studies addressing its life-history, home-range, habitat preferences and the effects of pollutants (Aparicio and De Sostoa, 1999, Aparicio, 2002, Figuerola et al., 2012, Grossman and De Sostoa, 1994), a knowledge gap remains on the impact of biotic variables such as the density of accompanying fish species or invertebrate predictors in its ecological niche.

Therefore, the objective of this study was: (1) to test the proficiency of the MLP Ensembles to model the ecological niche of freshwater fish species, and (2) to test whether biotic variables affect the distribution of redfin barbel. To achieve these aims using MLP Ensembles, two different models of redfin barbel were developed. The first considered only physical habitat variables, the second included biotic and physical habitat variables.

2 Materials and methods

2.1 Data collection

The study was conducted at the meso-scale in every summer, between 2003 and 2006. The study sites were located in the headwaters of the Ebron and Vallanca Rivers (Turia River tributaries), the Palancia River and the Villahermosa River (Mijares River Tributary) (Fig. 1) which approximately correspond to the southern limits of redfin barbel distribution (Perea et al., 2011). All the study sites were in unregulated streams and therefore a wide flow range was sampled (i.e. from 0.02 m^3/s to 1.84 m^3/s). For complete climatic description of the study area, see Alcaraz-Hernández et al. (2011) and Mouton et al. (2011).

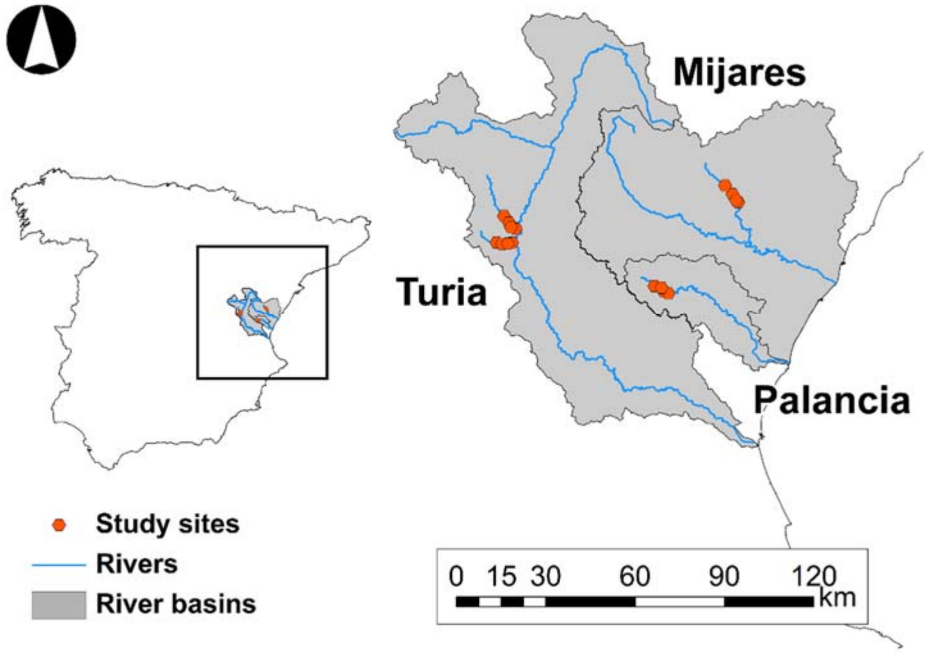


Fig. 1. Location of the target river basins in the Iberian Peninsula (left) and study sites in the Mijares, Palancia and Turia River basins.

2.1.1 Physical habitat survey

The physical habitat was assessed in every 300 m reach using an adaptation of the Basinwide Visual Estimation Technique (Dolloff et al., 1993). The approach stratifies the study site by HydroMorphological Units (hereafter called HMUs) classified as: pools, glides, riffles, and rapids (see Alcaraz-Hernández et al., 2011 for further details). Once an HMU was categorised, its physical attributes were recorded. They were, length, average width, obtained from three cross-sections corresponding to ¼, ½, and ¾ of the total length, mean depth (hereafter as depth), calculated from nine points corresponding to the measurements taken at each of the aforementioned cross-sections and the maximum depth, measured at the corresponding point. Percentage of shading over the channel, percentage of embeddedness, pieces of woody debris and percentage of the substrate types following a simplified classification from the American Geophysical Union (Martínez-Capel et al., 2009, Muñoz-Mas et al., 2012) were visually estimated and summarised in the substrate index (Mouton et al., 2011). In addition, the cover index (García de Jalón and Schmidt, 1995) was determined. This index characterises the available refuge due to caves, shading, substrate, submerged vegetation and water depth by assigning six scores from 0 (no refuge) to 5 (maximum score), and the weighted aggregation of these scores produces an index range from 0 to 10.

The river flow was gauged in at least one cross-section using an electromagnetic current meter (Valeport®), and flow velocity was calculated by dividing the flow by the average cross-section area. Elevation and slope were extracted from cartography in a geographic information system, whereas habitat variability was estimated with the Shannon-Weaver diversity index, taking into account the number of habitat types (i.e. number of pools, glides, riffles or rapids) from the visual stratification of each study site.

Table 1. Code, summary, units and description of the variables included in MLP Ensemble models.

Variable code	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Units	Description
River.Reach	1	4	8	8.108	12	16	-	Study site code
Year	2003	2004	2005	2005	2006	2006	-	Date
Meso.type	1	2	3	2.95	4	4	-	Mesohabitat type
Meso.diversity	0.2	0.62	0.7	0.68	0.76	0.99	-	Mesohabitat diversity
Length	8.6	19.1	24.36	26.92	31.5	54.7	m	Mean depth
Width	1.26	3.43	4.79	4.66	5.83	8.8	m	Maximum depth
Depth	0.04	0.22	0.32	0.35	0.46	0.79	m	Length
M.Depth	0.15	0.43	0.63	0.64	0.83	1.23	m	Width
Velocity	0.01	0.09	0.24	0.3	0.42	1.06	m/s	Mean flow velocity
Substrate	2.65	4.9	5.2	5.22	5.7	8	-	Substrate index
Embededness	0	0	15	29.35	50	100	%	% mud covering substrate
Cover	1	2.75	3.5	3.67	4.25	7.5	-	Cover index
Shadow	0	20	60	54.95	85	100	%	% shading
Wood.debries	0	0	0	0.01	0	0.16	pieces/m^2	Woody debris
Elevation	605	655	743	745.8	792	968	m	Elevation
Slope	0.01	0.01	0.01	0.02	0.02	0.04	m/m	Slope
D.redfin	0	0	0	2.61	2.66	31.22	ind./100 m^2	Density of *Barbus haasi*
D.b.trout	0	1.32	4.48	11.19	16.12	86.47	ind./100 m^2	Density of *Salmo trutta*
D.eel	0	0	0	0.9	0	20.74	ind./100 m^2	Density of *Anguilla anguilla*
D.r.trout	0	0	0	3.64	2.18	42.58	ind./100 m^2	Density of *Oncorhynchus mykiss*
D.cyprinids	0	0	0	12.31	11.09	198.3	ind./100 m^2	Cyprinids density
Inv.density	0	1930	4680	7910	9590	56010	ind./ m^2	Invertebrates density
Inv.richness	0	16	19	18.55	22	34	-	Invertebrates richness
Inv.diversity	0	0.34	0.41	0.39	0.46	0.56	-	Invertebrates diversity
Inv.biomass	0	0	0.07	0.39	0.27	6.02	g/m^2	Invertebrates biomass

2.1.2 Biological survey

The biological survey was undertaken by electrofishing, and all captured fish species were recorded. In each study site, one slow (i.e. pool or glide) and one fast (i.e. riffle or rapid) HMU were selected and surveyed (3-passes removal) after netting off the HMU. Due to a severe drought, some study sites were dry, resulting in 93 HMUs being sampled. Redfin barbel males are mature at approximately 45 mm, while females are mature at 100 mm (Aparicio, 2002), and therefore it was regarded as conservative to consider all specimens larger than 45 mm, resulting in a prevalence of 0.42. No size restrictions were imposed on the remaining fish species, and thus all the individuals were considered in the data analysis. Since the fish community varied across streams, the cyprinid species were grouped in a single variable (Table 2), and following previous studies fish densities were log(x+1) transformed (Brosse and Lek, 2000, Fukuda et al., 2011).

Table 2. Fish community in the four rivers. The cyprinid fish community varied across rivers and was summarized in a single variable.

Ebron	Vallanca	Palancia	Villahermosa
Salmo trutta	*Salmo trutta*	*Salmo trutta*	*Salmo trutta*
Oncorhynchus mykiss	*Oncorhynchus mykiss*	*Oncorhynchus mykiss*	*Oncorhynchus mykiss*
Babus haasi	*Babus haasi*	*Babus haasi*	*Babus haasi*
Luciobarbus guiraonis	*Luciobarbus guiraonis*	*Luciobarbus guiraonis*	*Luciobarbus guiraonis*
Anguilla anguilla	*Achondrostoma arcasii*	*Anguilla anguilla*	*Achondrostoma arcasii*
			Squalius valentinus
			Anguilla anguilla

Benthic invertebrates were collected with a Hess sampler (0.5 m^2) following the International Standard ISO 8265:1988, official version of the European Standard EN 29265 (January 1994). Samples were later identified to the lowest possible taxonomic level (predominantly at family level), sorted and counted to obtain the density of invertebrates. Specimens were dried in an oven at 65 ºC for 24 h and the dry residue was weighed to obtain invertebrate biomass. Finally, two additional predictors were derived: invertebrate richness (i.e. the sum of present taxa in each sample) and invertebrate diversity by applying the Shannon-Weaver diversity index based on the number of individuals per taxa at each sampled HMU (Table 1).

2.2 Models’ development

The physical habitat and biotic models were developed by means of MLP Ensembles (Hansen and Salamon, 1990).

The development of the optimal MLP Ensembles followed the overproduce-and-choose approach. This approach consists of the generation of an initial large pool of MLP candidate classifiers (overproduce) whereas the second phase is devoted to select the best performing subset of MLPs (choose). The choose phase was performed by means of the step-forward algorithm. Thus, starting from every MLP candidate classifier the best complementary MLP candidate is iteratively searched until no improvement in the Mean Squared Error (MSE) was achieved (Fig. 2 A).

To render parsimonious models, the optimal input variables’ subsets for both models were also selected by means of the step-forward algorithm. First, the best pair of input variables was determined by developing a MLP Ensemble for every uncorrelated pair following the aforementioned procedure and then this pair became the base for the following step forward variable selection. The algorithm continued until no more variables were available and the selected model was the one with the lowest number of variables and error. Finally, in order to rule out overfitting, we visually estimated differences between the distributions of the MSE based on the training and validation datasets of the selected MLPs (Fig. 2 B).

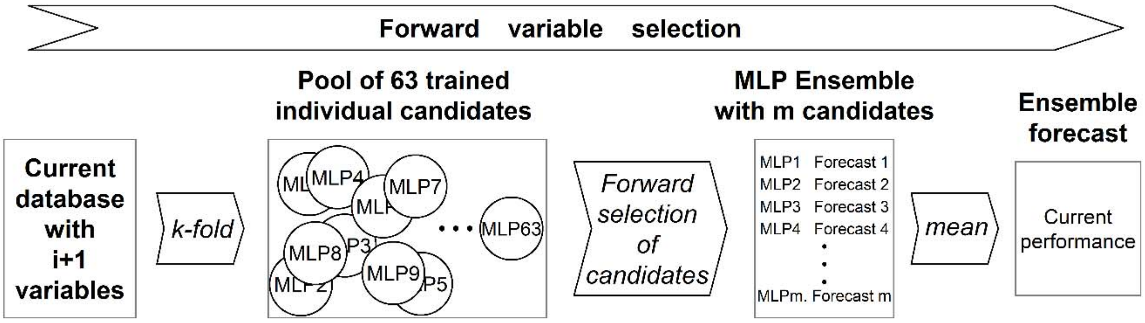

Forward variable selection
Pool of 63 trained individual candidates
MLP Ensemble with m candidates
Ensemble forecast
Current database with i+1 variables
k-fold
MLP4
MLP7
MLP63
MLP8
MLP9
P3
P5
Forward selection of candidates
MLP1 Forecast 1
MLP2 Forecast 2
MLP3 Forecast 3
MLP4 Forecast 4
MLPm. Forecast m
mean
Current performance

Fig. 2. Flowchart of the steps followed in the development of the physical habitat and biotic models.

2.2.1 Multilayer Perceptron Ensemble development

Building a MLP Ensemble involves training several individual models (MLPs) and combining them to produce aggregated predictions (Hansen and Salamon, 1990). The construction of the individual models (hereafter MLP candidates) was carried out in *R* (R Development Core Team, 2012) with the package *monmlp* which optimises the model weights using the non-linear minimisation (*nlm)* routine (Cannon, 2012). The activation functions were the hyperbolic tangent and the linear transformation, while the number of nodes was restricted to improve generalisation (Özesmi et al., 2006) following equation 1.

$$N_{nodes} = \max(\{1, \lfloor(number\ of\ variables\ +\ 1)/2\rfloor\})\ (1)$$

The aggregated forecast was determined by averaging the individual predictions of each selected model. Since model training depends on initial conditions, every MLP candidate was optimised five times with 500 iterations each.

Heterogeneity, or diversity, between MLP candidates is crucial because MLP Ensembles achieve better generalisation when models are complementary (Opitz, 1999). Several approaches allow for the construction of an MLP Ensemble with heterogeneous candidates, such as using different training datasets, architectures or learning methods (Brown et al., 2005). To increase heterogeneity among models, the database was divided in 63 different training and validation datasets corresponding to all possible combinations of 66 % of the cases for training, and 33 % for validation (i.e. following the *k-fold* approach). Consequently, 63 different MLP candidates were trained for every tested combination of input variables.

Originally the MLP Ensembles included all the developed models (Hansen and Salamon, 1990), but it was promptly demonstrated that active selection of the MLP candidates improved the final predictions (Opitz and Shavlik, 1996, Zhou et al., 2002). There are several methods to apply the overproduce-and-choose approach (Soares et al., 2013, Wang and Alhamdoosh, 2013, Yao and Xu, 2006, Zhou et al., 2002), but to our knowledge those sophisticated methods have not been coupled to a variable selection procedure. Consequently, we applied a step-forward selection of the MLP candidates which has been proved to perform similarly to more complex algorithms (Muñoz-Mas et al., 2014). Our step-forward selection was run starting from each of the 63 MLP candidates, searching for the best combination and stopping when no improvement was achieved. This was in contrast to the usual step-forward routine where the procedure would start from the best single model.

Since the optimal MLP Ensemble may not include all the MLP candidates, the observed performance could be affected by overfitting because the selected models may be trained only with some parts of the training database. Therefore the role of test data was twofold; first we applied an *a priori regularisation* method with the early stop regularisation (*sensu* Ludwig et al., 2014) by calculating the MSE on the validation dataset every 100 iterations of the *nlm* routine and then we visually estimated for each selected MLP differences between the distributions of the MSE based on the training and validation datasets of all the selected MLPs. In the case of dissimilar distributions, the number of nodes and the number of iterations ran between calculations of the MSE on the validation dataset were readjusted.

To allow for the comparison with previous studies that either included all trained networks (Palialexis et al., 2011) or based the model selection on a ranking of the individual performances, including the top MLPs (Watts and Worner, 2008), the MSE of the best MLP candidate and of the MLP Ensemble Complete (i.e. the one without any models' selection) and then for the best five, ten and fifteen models were calculated and compared with our optimal MLP Ensembles.

2.2.2 Variable selection

To identify the most important variables shaping the ecological niche, for both models an input variable subset was selected based on the step-forward procedure because it has proven computationally efficient and tends to result in relatively small input variables' subsets (May et al., 2011). In contrast to some other approaches (e.g. Generalised Additive Mixed Models, Lin and Zhang, 1999) the MLP Ensemble approach does not specifically allow for the consideration of spatial or temporal autocorrelation among training data. To rule out any influence of study site and sampling year, they were included as input variables (Table 1). Their absence on the ultimate models would indicate their irrelevance, thus corroborating the properness of the data packing. In addition, to render parsimonious models, instead of the usual step-forward procedure that discontinues when no improvement is achieved, the procedure was sustained until no more variables were available. The performance of the best model (MLP Ensemble) and the number of variables considered at every iteration were rescaled between 0 and 1 (1 being optimal), with the optimal MLP Ensemble being the one that maximised the sum of both criteria. The step-forward procedure may fail to consider variable interactions and may depend on the variable that was selected first. To overcome this limitation, one model was developed for each pairwise combination of variables. The best pair of variables was selected as the starting set of variables in the step-forward procedure. Additionally, during the entire process, neither correlated ($r^2 > 0.5$)

nor collinear (variable inflation factor; *vif* > 5) combinations of variables were considered. Since the input database was a combination of ordinal and continuous variables, the function *hetcor* in the package *polycor* (Fox, 2010) was used to calculate the variables' correlation (Appendix A).

2.3 Partial dependence plots and uncertainty analysis

Model reliability and transparency is of major concern for ecological modelling (Austin, 2007, Guisan and Thuiller, 2005, Özesmi et al., 2006) and is fundamental when models are used with exploratory purposes. Therefore, to graphically characterise the relationship between the input variables and the predicted densities obtained by the optimal MLP Ensembles, partial dependence plots (Friedman, 2001) implemented in the package *randomForests* (Liaw and Wiener, 2002) were developed.

The importance of dealing with uncertainty has been stressed as a key challenge in ecological modelling (Larocque et al., 2011). Consequently, partial dependence plots were developed also for every model in the optimal MLP Ensemble, and the function *densregion.normal* in the package *denstrip* (Jackson, 2008) was used to visually inspect the uncertainty associated to the MLP aggregation in comparison with the input variable distribution.

3 Results

3.1 Training results

The optimal physical habitat model included five variables (three nodes): elevation, embeddedness, depth, slope and cover (Fig. 3) with a maximum correlation of 0.33 and a variable inflation factor of 1.41. The optimal biotic model also included five variables (three nodes): density of eel, cyprinids' density, width, invertebrates' density and cover (Fig. 3) with a maximum correlation of 0.38 and variable inflation factor of 1.51. In addition, the spatiotemporal correlation was considered negligible since study site and sampling year were not selected as inputs in the ultimate models (i.e. the physical habitat and the biotic models).

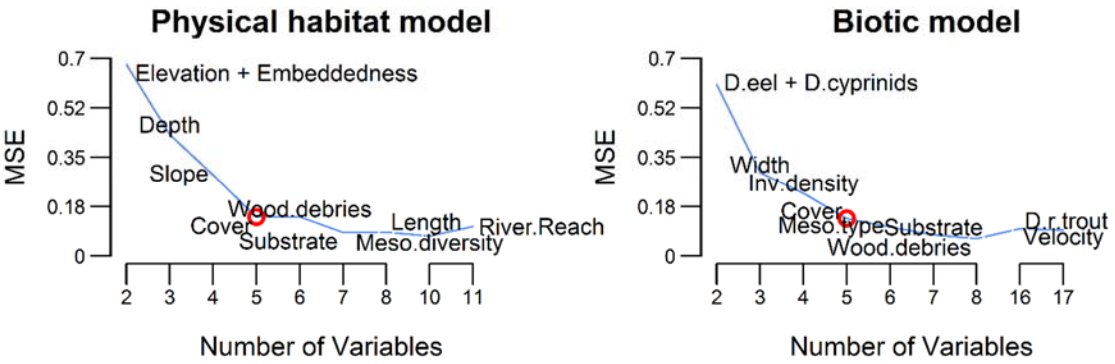


Fig. 3. Sequence of the variable selection during the step-forward procedure to develop the MLP Ensemble (from left to right). The plots show the Mean Squared Error, MSE, in function of the number of variables. The circle indicates the selection of the optimal model.

Although the two models showed similar performance (i.e. similar values of the Mean Squared Error, MSE) and the relatively large amount of zeros in the training dataset slightly biased their outputs, the biotic model slightly outperformed the physical habitat model with MSE of 0.12 and 0.13 respectively (Fig. 4 & Table 3). A complete description of the calculated MSE in every iteration is given in Appendix B.

The physical habitat model selected 15 MLPs. Consequently, 15 training datasets and 15 validation datasets were involved in its development. Cross-evaluation (i.e. the evaluation of every training and validation dataset with every selected MLP candidate) showed that MSEs were distributed equally for the training and the validation datasets (Fig. 5), and therefore we considered the physical habitat model not overfitted. The biotic model selected eight MLPs, with eight training and validation datasets involved in the development of the selected MLPs. Likewise, the distribution of the training and validation MSE clearly overlapped, and therefore the biotic model was also considered not overfitted. The training datasets of the MLP candidates selected within the optimal MLP Ensembles are described in Appendix C.

The best MLP candidate, the MLP Ensemble Complete (i.e. considering all sixty three MLPs) and the top five, top ten and top fifteen MLPs yielded higher MSEs than the optimal MLP Ensembles. The highest difference appeared between the MLP Ensemble Complete and the optimal physical habitat model (Table 3).

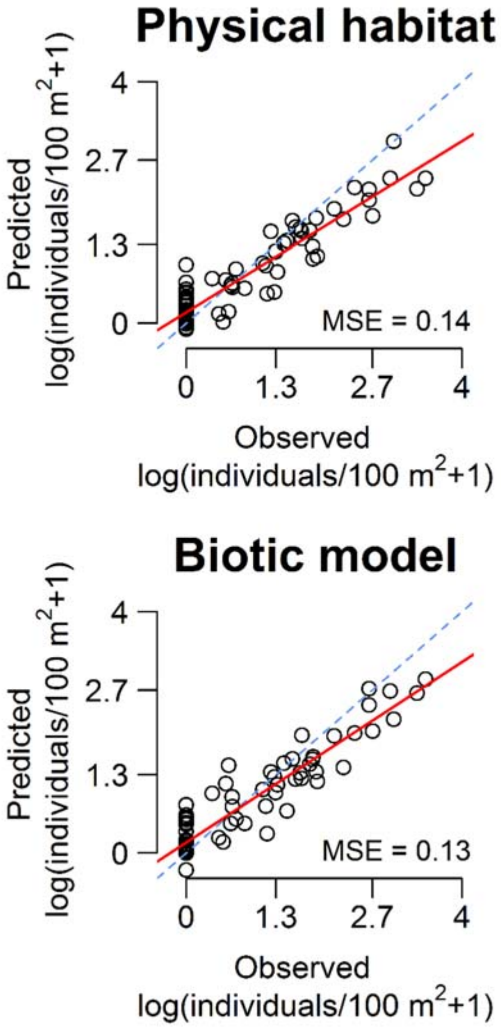


Fig. 4. Relation between the observed and predicted values of the optimal MLPs. The optimal physical habitat and biotic models show transformed output.

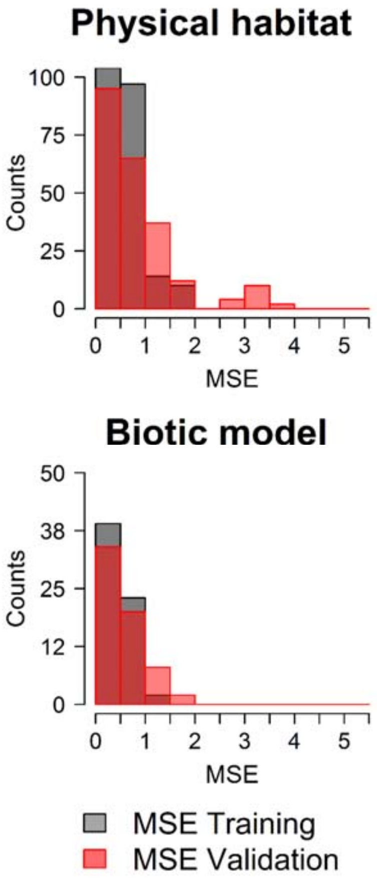


Fig. 5. Frequency analysis of the Mean Squared Error (MSE) of the selected MLP candidates based on the corresponding training and validation datasets.

Table 3. Mean Squared Error (MSE) of the best single MLP, the MLP Ensemble and the MLP Ensemble Complete, (without candidates selection) and MSE selecting the top 5, top 10 and top 15 MLP Candidates. The amount of considered networks appears in brackets.

Model	Physical habitat model	Biotic model
Best Single MLP	0.35 (1)	0.3 (1)
MLP Ensemble	0.13 (9)	0.12 (12)
MLP Ensemble Complete	2422.82 (63)	26.15 (63)
Top MLP - 1 to 5	0.19 (5)	0.18 (5)
Top MLP - 1 to 10	0.18 (10)	0.16 (10)
Top MLP - 1 to 15	0.19 (15)	0.17 (15)

3.2 Partial dependence plots - physical habitat model

The optimal physical habitat model showed a unimodal response between redfin barbel density and elevation, with a maximum density at 738 m above sea level. Embeddedness showed an almost flat trend but an exponential increase from 75 % onwards. Depth showed a steep positive linear trend, thus suggesting the major impact among the selected variables, whereas slope and cover were negatively and almost linearly related to redfin barbel density. As expected, uncertainty was higher at the extremes of the variables' distributions and therefore trends at these extreme values could be unreliable (Fig. 6).

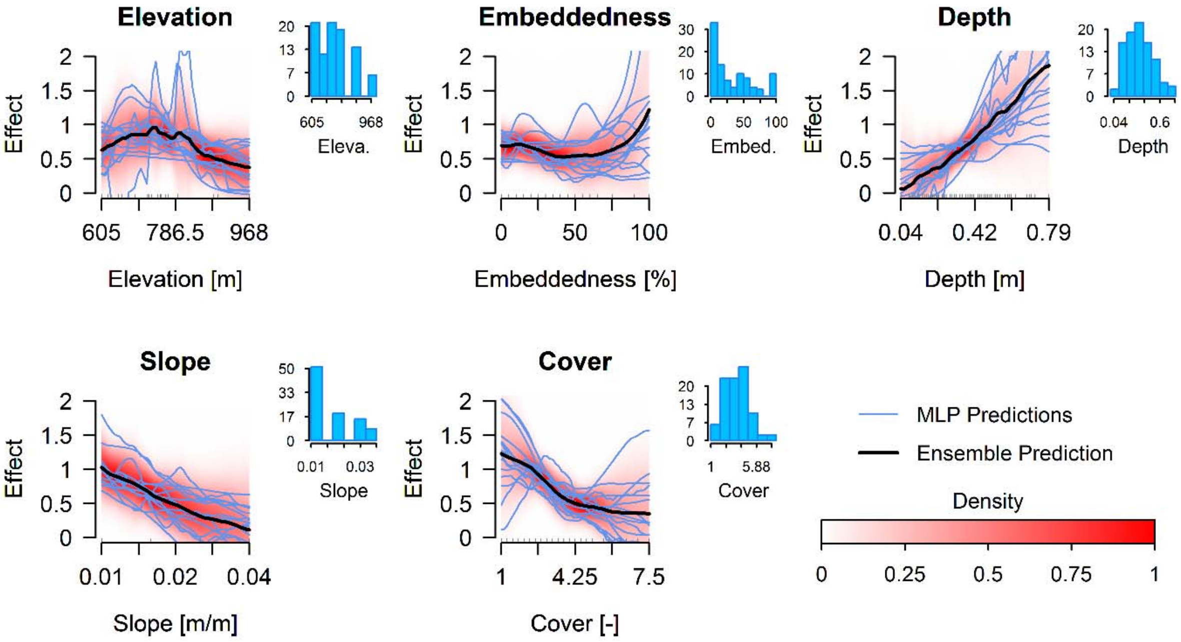


Fig. 6. Partial dependence plots of the physical habitat model (black line). Light lines (Blue lines on the e-version) correspond to the partial dependence plot of every selected MLP candidate. Faded background corresponds to uncertainty analysis based on mean and standard deviation of predictions, and the darker the colour the smaller the uncertainty.

3.3 Partial dependence plots - Biotic model

The optimal biotic model showed a positive linear relation between eel and redfin barbel densities. Cyprinids' density presented a unimodal response with the peak around 4 individuals/100 m^2. Width showed an almost positive linear influence on redfin barbel density whereas invertebrates' density presented a unimodal response inflecting at 21718 individuals/ m^2. Likewise the physical habitat model cover presented a linear trend but with smaller uncertainty and slope. Uncertainty was higher than in the physical habitat model although it presented a similar pattern with the extreme values being more uncertain than the central part of the input variables' distributions (Fig. 7).

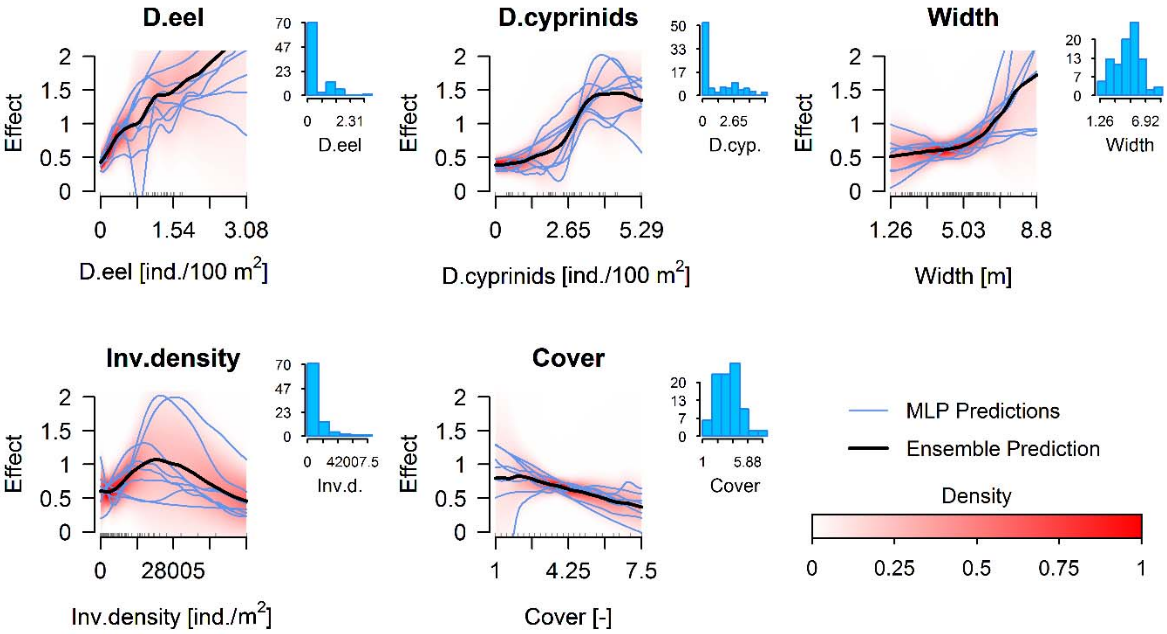


Fig. 7. Partial dependence plots of the biotic model (black line). Light lines (Blue lines on the e-version) correspond to the partial dependence plot of every selected MLP candidate. Faded background corresponds to uncertainty analysis based on mean and standard deviation of predictions, and the darker the colour the smaller the uncertainty.

4 Discussion

4.1 MLP Ensemble development

Our results indicated that the MLP Ensemble paradigm can be considered proficient to model the ecological niche of freshwater fish species, in line with previous studies that modelled fish density with neural networks (Baran et al., 1996, Brosse and Lek, 2000, Laffaille et al., 2003). The presented optimal models also outperformed any single MLP, which agrees with previous research (Palialexis et al., 2011). We also demonstrated that active selection of MLP candidates to create an optimal subset can further improve MLP Ensembles' performance. This is consistent with Zhou's *et al*. (2002) statement that "*many could be better than all*". Moreover, our candidates' selection approach resulted in a better performance in contrast with the selection approach based on the individual performance (i.e. top five, ten and fifteen). We recommend this procedure in contrast to previous studies that selected the best subset based on the individual performance of the MLPs (Watts and Worner, 2008). However, our step-forward process is determined by the first selected model, and despite the fact that the procedure started from every single neural network, the possibility to get stuck at a local minimum exists. Therefore, untested combinations of models could outperform those obtained by the step-forward algorithm.

Genetic algorithms may overcome the aforementioned constraints (Soares et al., 2013, Wang and Alhamdoosh, 2013). There are successful applications of genetic algorithms in variable selection procedures (May et al., 2011, Olden et al., 2008) and also within the selection of optimal MLP Ensembles (Soares et al., 2013, Wang and Alhamdoosh, 2013). Although the use of genetic algorithms for both variable selection and MLP candidates' selection could exponentially increase the computation effort, this approach is certainly promising and should be the subject of future research.

The effect of the relatively large amount of zeros in the training dataset was a remarkable issue and slightly biasing models' outputs. There are specific techniques in count data modelling to deal with excess of zeros with either parametric (Mullahy, 1986, Lambert, 1992) and non-parametric (Liu and Chan, 2010) responses. Certainly the comparison of the capability of MLP Ensembles and these techniques in ecological modelling would be of interest, although these techniques by definition do not easily account for variable interactions and thus do not easily assure better performance. Nevertheless, our results were considered satisfactory since they provided an acceptable balance between model complexity, performance and computational effort, and they were devoted to explore rather than to predict fish density in further analysis.

4.2 Ecological relevance of the physical habitat model

The optimal physical habitat model included five variables: elevation, slope, depth, embeddedness and cover. Elevation is broadly accepted as a proximal predictor of temperature (Elith and Leathwick, 2009), and consequently we considered that it may reflect the effect of climate on redfin barbel distribution. Similarly, the negative trend of redfin barbell density versus slope agrees with its preference for middle-to-upper stream reaches of mountainous rivers (Perea et al., 2011). The positive impact of depth corresponds with previous studies that considered the redfin barbel a pool dweller (Aparicio and De Sostoa, 1999). Our results may also suggest the importance of backwaters or stagnated areas as resting habitats. This could also explain the positive relationship between redfin barbell density and embeddedness, since pool substrates are generally more embedded. Despite the negative relationship between cover and redfin barbel density in our study, some authors classified the redfin barbel as a cover-oriented fish (Grossman and De Sostoa, 1994). Aparicio (2002) reported the active use of cover in an ephemeral river but related its use to the absence of deep pools in this specific river, rather than to a redfin barbel preference for cover. Our study suggests that in more complex river systems with well-developed pool-riffle patterns, redfin barbel may tend to avoid excessive cover complexity.

4.3 Ecological relevance of the biotic model

In contrast to previous studies (Vezza et al., 2015, Watts and Worner, 2008) where the combination of physical habitat and biotic variables outperformed the model developed only with physical habitat variables, our biotic model did not significantly perform better than the physical habitat model. Moreover, uncertainty was higher in the biotic model, which underlines the previously reported complexity related to the assessment of biotic interactions (Leathwick and Austin, 2001). Although larger datasets may reduce this uncertainty, our study nevertheless suggests interesting and plausible relationships. In addition, the observed associations between biotic variables and red fin density agreed with the ecological gradient theory because responses were quasi-linear or unimodal (Austin, 2007).

The biotic model demonstrated a positive association between redfin barbel and European eel, which confirms the work of Laffaille et al. (2003). They modelled eel habitat suitability in a small coastal catchment with a single MLP. Eels were more abundant in deep and low flow shaded areas without aquatic vegetation (Laffaille et al., 2003). Such a pattern broadly concurs with the requirements of redfin barbel. Also the suggested relationship between redfin barbel and cyprinids corresponds with previous studies on the Iberian Peninsula that reported the presence of multi-species shoals as well as an overlap in microhabitat use (Martínez-Capel et al., 2009). Fish schooling benefits include the enhancement of hydrodynamics and the protection against predators (Landa, 1998). Moreover, similar positive interactions with cyprinid species have been reported for Iberian chub (*Squalius pyrenaicus*, Günther, 1868) and eastern Iberian barbel (*Luciobarbus guiraonis*, Steindachner, 1866) (Vezza et al., 2015). This indicates that restoration actions focused on redfin barbel could also result in habitat enhancement for other cyprinid species.

Although previous work positively correlated invertebrate density to fish density (Mas-Martí et al., 2010), our results show a maximal redfin barbel density at 21718 individuals/m^2. This could be related to food availability. However, the preferred prey invertebrates of redfin barbel (i.e. *Chironomidae*, *Ephemeroptera* and *Trichoptera* following Miranda et al., 2005) were strongly correlated with the invertebrate density applied in our model. Therefore, we attributed the avoidance of the higher invertebrate densities to a habitat correlation. Previous studies suggested that in Mediterranean rivers, with very unstable climatic conditions, riffles tend to host higher invertebrate density than pools (Bonada et al., 2006). Therefore the decrement of the partial dependence plot could be showing the necessity for larger depth rather than a preference for intermediate invertebrate densities. Nevertheless, the discrepancy between our results and the literature can also be a consequence of the applied model complexity (e.g. linear *vs.* non-linear models), and the impact of model complexity should be thoroughly analysed in further studies modelling the relationship between invertebrate and fish density. The biotic model also selected two physical habitat variables, cover index and width, and the relationship with cover index being similar to that shown in the physical habitat model. The positive association between redfin barbel density and width may be attributed to the negative correlation between width and elevation, in line with the aforementioned redfin barbel preference for middle-to-upper stream reaches, but with a slightly different response because the study encompassed four different rivers.

Although the biotic model appears to suggest interactions between redfin barbel and other species, significant positive or negative correlations between species does not imply a causative effect (Wisz et al., 2013). A simple correlation does not mandatorily correspond to any species interaction, neither mutualism nor facilitation, and therefore further research should clarify the true impact of species interactions. Furthermore, changes in the habitat available may result in a substantial increment in the competition between species (Wisz et al., 2013). The cyprinids density partial dependence plot showed a decrement at the tail of the curve. Therefore, in spite of being uncertain, it could suggest that, under different habitat conditions than those in our study, the positive interaction with cyprinid fish species may become habitat competition, thus emphasising the necessity of close monitoring in the near future to avoid ecological loss.

4.4 Model uncertainty

Relatively few studies address uncertainty in ecological modelling and its effects on model predictions and decision making (Elith and Leathwick, 2009). In accordance with previous studies (Peters et al., 2009), the largest uncertainty tended to appear in the regions of the input variables that were poorly represented in the training database. In contrast to the high uncertainty demonstrated by the different MLP candidate predictions, the optimal MLP Ensembles produced sound and smoother partial dependence plots that allowed general trends to be derived from a wide range of model outputs. This has been stimulated by three approaches applied in our study. First, the bias and variance dilemma (Geman et al., 1992) was addressed by limiting model complexity (i.e. limiting the number of nodes and variables), leading to less complex models than in previous studies (Dedecker et al., 2004, Lek et al., 1996). Second, the early stop regularisation considered the errors committed on the training and validation datasets (Ludwig et al., 2014). Third and most importantly, we assessed overfitting by checking whether species responses to environmental variables were consistent with the ecological gradient theory (Austin, 2007). Inconsistent model results would have suggested a more restrictive modelling approach by limiting model complexity or adjusting the early stop parameters. Uncertainty could also arise when samples from different periods are combined, since fish density is a density-dependent phenomenon (Mas-Martí et al., 2010); however, the sampling year was not selected as an important variable in the optimal MLP Ensembles. Consequently, the results from our study suggest that temporal packing can be considered admissible when focusing on a short time span.

4.5 Conclusions

The MLP Ensembles have demonstrated satisfactory performance and generated predictions in line with the ecological gradient theory. As such, the models provide a better insight into the ecological niche of redfin barbel by complementing previous studies. We not only demonstrated that MLP Ensembles may outperform single MLPs, but also that the active selection of MLP candidates to create an optimal subset of models can further improve model performance. The physical habitat model confirmed the redfin barbel preference for middle-to-upper river segments, but not in higher and steeper reaches. The importance of depth confirms that redfin barbel prefer pool-type habitats, which emphasises the vulnerability of the species to reduced flows. Although the biotic model showed higher uncertainty, it suggested that redfin barbel, European eel and the considered cyprinid species have similar habitat requirements. Due to its high predictive performance and its ability to deal with model uncertainty, the MLP Ensemble paradigm is a promising tool for ecological modelling or habitat suitability prediction to assess environmental flows.

Acknowledgements

This study was funded by the Spanish Ministry of Economy and Competitiveness with the project SCARCE (Consolider-Ingenio 2010 CSD2009-00065) and the Universitat Politècnica de València, through the project UPPTE/2012/294 (PAID-06-12). Additionally, the authors would like to thank the help of the Conselleria de Territori i Vivenda (Generalitat Valenciana) and the Confederación Hidrográfica del Júcar (Spanish government) which provided environmental data. The authors are indebted to all the colleagues who collaborated in the field data collection and the text adequacy; without their help this paper would have not been possible. Last but not least, the authors would like to specifically thank E. Aparicio and A.J. Cannon, the former because he selflessly provided

the bibliography about the redfin barbel and the latter because he patiently explained the ‘ins and outs’ of the *monmlp* package.

5 References

Alcaraz-Hernández, J.D., Martínez-Capel, F., Peredo, M. and Hernández-Mascarell, A., 2011. Mesohabitat heterogeneity in four mediterranean streams of the Jucar river basin (Eastern Spain). Limnetica 30 (2), 15-363.

Aparicio, E., 2002. Ecologia del barb cua-roig (Barbus haasi) i avaluació del seu estat de conservació a Catalunya. Programa de Doctorat de Biologia Animal I - Zoologia - Bienni 1991-1993, 173 (Catalan).

Aparicio, E. and De Sostoa, A., 1999. Pattern of movements of adult Barbus haasi in a small Mediterranean stream. J. Fish Biol. 55 (5), 1086-1095. http://dx.doi.org/10.1006/jfbi.1999.1109.

Araújo, M.B. and New, M., 2007. Ensemble forecasting of species distributions. Trends Ecol. Evol. 22 (1), 42-47. http://dx.doi.org/10.1016/j.tree.2006.09.010.

Austin, M., 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecol. Model. 200 (1-2), 1-19. http://dx.doi.org/10.1016/j.ecolmodel.2006.07.005.

Baran, P., Lek, S., Delacoste, M. and Belaud, A., 1996. Stochastic models that predict trout population density or biomass on a mesohabitat scale. Hydrobiologia 337 (1-3), 1-9. http://dx.doi.org/10.1007/BF00028502.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., et al, 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. Ecol. Model. 222 (11), 1810-1819. http://dx.doi.org/10.1016/j.ecolmodel.2011.02.011.

Bianco, P.G., 1998. Diversity of Barbinae fishes in southern Europe with description of a new genus and a new species (Cyprinidae). Ital. J. Zool. 65 (Suppl. 1), 125-136. http://dx.doi.org/10.1080/11250009809386804.

Bonada, N., Rieradevall, M., Prat, N. and Resh, V.H., 2006. Benthic macroinvertebrate assemblages and macrohabitat connectivity in Mediterranean-climate streams of northern California. J. N. Am. Benthol. Soc. 25 (1), 32-43. http://dx.doi.org/10.1899/0887-3593(2006)25[32:bmaamc]2.0.co;2.

Boulangeat, I., Gravel, D. and Thuiller, W., 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. Ecol. Lett. 15 (6), 584-593. http://dx.doi.org/10.1111/j.1461-0248.2012.01772.x.

Brosse, S. and Lek, S., 2000. Modelling roach (Rutilus rutilus) microhabitat using linear and nonlinear techniques. Freshwater Biol. 44 (3), 441-452. http://dx.doi.org/10.1046/j.1365-2427.2000.00580.x.

Brown, G., Wyatt, J., Harris, R. and Yao, X., 2005. Diversity creation methods: A survey and categorisation. Inf. Fusion 6 (1), 5-20. http://dx.doi.org/10.1016/j.inffus.2004.04.004.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. and Grenouillet, G., 2010. Uncertainty in ensemble forecasting of species distribution. Glob. Change Biol. 16 (4), 1145-1157. http://dx.doi.org/10.1111/j.1365-2486.2009.02000.x.

Cannon, A.J., 2012. monmlp: Monotone multi-layer perceptron neural network. R package version 1.1.2.

Crisci, C., Ghattas, B. and Perera, G., 2012. A review of supervised machine learning algorithms and their applications to ecological data. Ecol. Model. 240 (0), 113-122. http://dx.doi.org/10.1016/j.ecolmodel.2012.03.001.

Dedecker, A.P., Goethals, P.L.M., Gabriels, W. and De Pauw, N., 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecol. Model. 174 (1-2), 161-173. http://dx.doi.org/10.1016/j.ecolmodel.2004.01.003.

Dolloff, C.A., Hankin, D.G. and Reeves, G.H., 1993. Basinwide Estimation of Habitat and Fish Populations in Streams Gen. Tech. Rep. SE-83., Southeastern Forest Experiment Station, Asheville, North Carolina (USA).

Elith, J. and Leathwick, J.R., 2009. Species distribution models: Ecological explanation and prediction across space and time. Ann. Rev. Ecol. Evol. Syst. 40 677-697. http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159.

Ferreira, T., Oliveira, J., Caiola, N., de Sostoa, A., Casals, F., Cortes, R., et al, 2007. Ecological traits of fish assemblages from Mediterranean Europe and their responses to human disturbance. Fisheries Manag. Ecol. 14 (6), 473-481. http://dx.doi.org/10.1111/j.1365-2400.2007.00584.x.

Figuerola, B., Maceda-Veiga, A. and de Sostoa, A., 2012. Assessing the effects of sewage effluents in a Mediterranean creek: Fish population features and biotic indices. Hydrobiologia 694 (1), 75-86. http://dx.doi.org/10.1007/s10750-012-1132-y.

Fox, J., 2010. polycor: Polychoric and Polyserial Correlations. R package version 0.7-8.

Freyhof, J. and Brooks, E., 2011. European Red List of Freshwater Fishes Luxembourg (Luxembourg).

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Stat. 29 (5), 1189-1232. http://dx.doi.org/10.1214/aos/1013203451.

Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J. and Mouton, A.M., 2013. Habitat prediction and knowledge extraction for spawning European grayling (*Thymallus thymallus* L.)

using a broad range of species distribution models. Environ. Modell. Softw. 47 1-6. http://dx.doi.org/10.1016/j.envsoft.2013.04.005.

Fukuda, S., Mouton, A.M. and De Baets, B., 2011. Abundance versus presence/absence data for modelling fish habitat preference with a genetic Takagi-Sugeno fuzzy system. Environ. Monit. Assess. 184 (10), 6159–6171. http://dx.doi.org/10.1007/s10661-011-2410-2.

Fukuda, S., De Baets, B., Mouton, A.M., Waegeman, W., Nakajima, J., Mukai, T., et al, 2011. Effect of model formulation on the optimization of a genetic Takagi–Sugeno fuzzy system for fish habitat suitability evaluation. Ecol. Model. 222 (8), 1401-1413. http://dx.doi.org/10.1016/j.ecolmodel.2011.01.023.

García de Jalón, D. and Schmidt, G., 1995. Manual práctico para la gestión sostenible de la pesca fluvial. Madrid, (Spain) (Spanish).

Geman, S., Bienenstock, E. and Doursat, R., 1992. Neural networks and the bias/variance dilemma. Neural Comput. 4 (1), 1–58. http://dx.doi.org/10.1162/neco.1992.4.1.1.

Gevrey, M., Dimopoulos, I. and Lek, S., 2006. Two-way interaction of input variables in the sensitivity analysis of neural network models. Ecol. Model. 195 (1-2), 43-50. http://dx.doi.org/10.1016/j.ecolmodel.2005.11.008.

Grossman, G.D. and De Sostoa, A., 1994. Microhabit use by fish in the upper Rio Matarrana, Spain, 1984-1987. Ecol. Freshwat. Fish 3 (4), 141-152. http://dx.doi.org/10.1111/j.1600-0633.1994.tb00016.x.

Guisan, A. and Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009. http://dx.doi.org/10.1111/j.1461-0248.2005.00792.x.

Hansen, L.K. and Salamon, P., 1990. Neural network ensembles. IEEE T. Pattern Anal. 12 (10), 993-1001. http://dx.doi.org/10.1109/34.58871.

Hopkins II, R.L. and Burr, B.M., 2009. Modeling freshwater fish distributions using multiscale landscape data: A case study of six narrow range endemics. Ecol. Model. 220 (17), 2024-2034. http://dx.doi.org/10.1016/j.ecolmodel.2009.04.027.

Mullahy, J., 1986. Specification and testing of some modified count data models. J. Econom. 33 (3), 341-365. http://dx.doi.org/10.1016/0304-4076(86)90002-3.

Jackson, C.H., 2008. Displaying uncertainty with shading. Am. Stat. 4 (62), 340-347. http://dx.doi.org/10.1198/000313008X370843.

Kemp, S.J., Zaradic, P. and Hansen, F., 2007. An approach for determining relative input parameter importance and significance in artificial neural networks. Ecol. Model. 204 (3–4), 326-334. http://dx.doi.org/10.1016/j.ecolmodel.2007.01.009.

Laffaille, P., Feunteun, E., Baisez, A., Robinet, T., Acou, A., Legault, A., et al, 2003. Spatial organisation of European eel (Anguilla anguilla L.) in a small catchment. Ecol. Freshwat. Fish 12 (4), 254-264. http://dx.doi.org/10.1046/j.1600-0633.2003.00021.x.

Lambert, D., 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics. 34, 11-14. http://dx.doi.org/10.2307/1269547.

Landa, J.T., 1998. Bioeconomics of schooling fishes: Selfish fish, quasi-free riders, and other fishy tales. Environ. Biol. Fish. 53 (4), 353-364. http://dx.doi.org/10.1023/A:1007414603324.

Larocque, G.R., Mailly, D., Yue, T.-., Anand, M., Peng, C., Kazanci, C., et al, 2011. Common challenges for ecological modelling: Synthesis of facilitated discussions held at the symposia organized for the 2009 conference of the International Society for Ecological Modelling in Quebec City, Canada, (October 6-9, 2009). Ecol. Model. 222 (14), 2456-2468. http://dx.doi.org/10.1016/j.ecolmodel.2010.12.017.

Leathwick, J.R. and Austin, M.P., 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. Ecology 82 (9), 2560-2573. http://dx.doi.org/10.2307/2679936.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. and Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90 (1), 39-52. http://dx.doi.org/10.1016/0304-3800(95)00142-5.

Liaw, A. and Wiener, M., 2002. Classification and Regression by randomForest. R News 3 (2), 18-22.

Lin, X. and Zhang, D., 1999. Inference in generalized additive mixed models by using smoothing splines. J. R. Stat. Soc. Ser. B Stat. Methodol. 61 (2), 381-400. http://dx.doi.org/10.1111/1467-9868.00183.

Liu, H. and Chan, K.-., 2010. Introducing COZIGAM: An R package for unconstrained and constrained zero-inflated generalized additive model analysis. J. Stat. Software 35 (11), 1-26.

Ludwig, O., Nunes, U. and Araujo, R., 2014. Eigenvalue decay: A new method for neural network regularization. Neurocomputing 124 33-42. http://dx.doi.org/10.1016/j.neucom.2013.08.005.

Martínez-Capel, F., García De Jalón, D., Werenitzky, D., Baeza, D. and Rodilla-Alamá, M., 2009. Microhabitat use by three endemic Iberian cyprinids in Mediterranean rivers (Tagus River Basin, Spain). Fisheries Manag. Ecol. 16 (1), 52-60. http://dx.doi.org/10.1111/j.1365-2400.2008.00645.x.

Mas-Martí, E., García-Berthou, E., Sabater, S., Tomanova, S. and Muñoz, I., 2010. Comparing fish assemblages and trophic ecology of permanent and intermittent reaches in a Mediterranean stream. Hydrobiologia 657 (1), 167-180. http://dx.doi.org/10.1007/s10750-010-0292-x.

May, R., Dandy, G. and Maier, H., 2011. Review of Input Variable Selection Methods for Artificial Neural Networks. In: Suzuki, K.(ed.), Artificial Neural Networks - Methodological Advances and Biomedical Applications. InTech., pp. 362.

Miranda, R., Díez-León, M. and Escala, M.C., 2005. Length relationships of cyprinid prey in diet analysis of Eurasian otter Lutra lutra in Mediterranean habitats. Folia Zool. 54 (4), 443-447.

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 (5), 615–622. http://dx.doi.org/10.1016/j.envsoft.2010.12.001.

Mouton, A.M., De Baets, B. and Goethals, P.L.M., 2010. Ecological relevance of performance criteria for species distribution models. Ecol. Model. 221 (16), 1995-2002. http://dx.doi.org/10.1016/j.ecolmodel.2010.04.017.

Muñoz-Mas, R., Alcaraz-Hernández, J.D. and Martínez-Capel, F., 2014. Multilayer Perceptron Ensembles (MLP Ensembles) in modelling microhabitat suitability for freshwater fish. XVII Congreso Español sobre Tecnologías y Lógica Fuzzy (ESTYLF 2014), Zaragoza (Spain), 609-614.

Muñoz-Mas, R., Martínez-Capel, F., Schneider, M. and Mouton, A.M., 2012. Assessment of brown trout habitat suitability in the Jucar River Basin (SPAIN): Comparison of data-driven approaches with fuzzy-logic models and univariate suitability curves. Sci. Total Environ. 440 123-131. http://dx.doi.org/10.1016/j.scitotenv.2012.07.074.

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 (1–2), 135–150. http://dx.doi.org/10.1016/s0304-3800(02)00064-9.

Olden, J.D., Lawler, J.J. and Poff, N.L., 2008. Machine learning methods without tears: A primer for ecologists. Q. Rev. Biol. 83 (2), 171-193. http://dx.doi.org/10.1086/587826.

Opitz, D.W., 1999. Feature selection for ensembles. Proceedings of the 1999 16th National Conference on Artificial Intelligence (AAAI-99), 11th Innovative Applications of Artificial Intelligence Conference (IAAI-99), Orlando, FL, (USA), 379-384.

Opitz, D.W. and Shavlik, J.W., 1996. Actively Searching for an Effective Neural Network Ensemble. Connect. Sci. 8 (3-4), 337-353. http://dx.doi.org/10.1080/095400996116802.

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 (1-2), 83-93. http://dx.doi.org/10.1016/j.ecolmodel.2005.11.012.

Palialexis, A., Georgakarakos, S., Karakassis, I., Lika, K. and Valavanis, V.D., 2011. Fish distribution predictions from different points of view: Comparing associative neural networks, geostatistics and regression models. Hydrobiologia 670 (1), 165-188. http://dx.doi.org/10.1007/s10750-011-0676-6.

Park, Y. and Chon, T., 2007. Biologically-inspired machine learning implemented to ecological informatics. Ecol. Model. 203 (1–2), 1-7. http://dx.doi.org/10.1016/j.ecolmodel.2006.05.039.

Perea, S., Garzón, P., González, J.L., Almada, V.C., Pereira, A. and Doadrio, I., 2011. New distribution data on Spanish autochthonous species of freshwater fish. Graellsia 67 (1), 91-102. http://dx.doi.org/10.3989/graellsia.2011.v67.032.

Peters, J., Verhoest, N.E.C., Samson, R., Van Meirvenne, M., Cockx, L. and De Baets, B., 2009. Uncertainty propagation in vegetation distribution models based on ensemble classifiers. Ecol. Model. 220 (6), 791-804. http://dx.doi.org/10.1016/j.ecolmodel.2008.12.022.

R Development Core Team, 2012. R: A language and environment for statistical computing.

Soares, S., Antunes, C.H. and Araújo, R., 2013. Comparison of a genetic algorithm and simulated annealing for automatic neural network ensemble development. Neurocomputing 121 498-511. http://dx.doi.org/10.1016/j.neucom.2013.05.024.

Soberón, J., 2007. Grinnellian and Eltonian niches and geographic distributions of species. Ecol. Lett. 10 (12), 1115-1123. http://dx.doi.org/10.1111/j.1461-0248.2007.01107.x.

Vezza, P., Muñoz-Mas, R., Martinez-Capel, F. and Mouton, A., 2015. Random forests to evaluate biotic interactions in fish distribution models. Environ. Model. Softw. 67 173-183. http://dx.doi.org/10.1016/j.envsoft.2015.01.005.

Wang, D. and Alhamdoosh, M., 2013. Evolutionary extreme learning machine ensembles with size control. Neurocomputing 102 98-110. http://dx.doi.org/10.1016/j.neucom.2011.12.046.

Watts, M.J. and Worner, S.P., 2008. Comparing ensemble and cascaded neural networks that combine biotic and abiotic variables to predict insect species distribution. Ecol. Inform. 3 (6), 354-366. http://dx.doi.org/10.1016/j.ecoinf.2008.08.003.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., et al, 2013. The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. Biol. Rev. 88 (1), 15-30. http://dx.doi.org/10.1111/j.1469-185X.2012.00235.x.

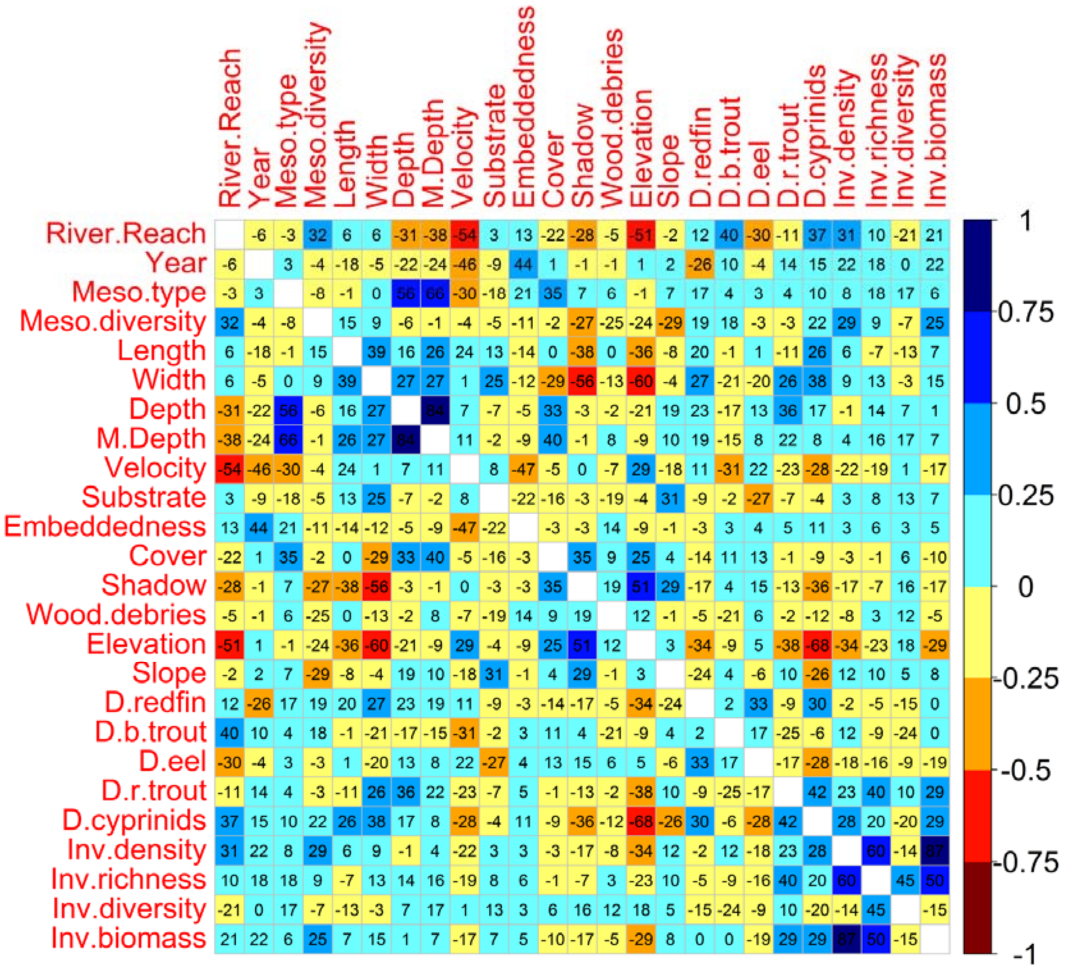
Yao, X. and Xu, Y., 2006. Recent advances in evolutionary computation. J. Comput. Sci. Technol. 21 (1), 1-18. http://dx.doi.org/10.1007/s11390-006-0001-4.

Zhou, Z.H., Wu, J. and Tang, W., 2002. Ensembling neural networks: Many could be better than all. Artif. Intell. 137 (1-2), 239-263. http://dx.doi.org/10.1016/s0004-3702(02)00190-x.

Appendix A1

1 Input variables correlation

The input database was a combination of ordinal and continuous variables then the function *hetcor* in the package *polycor* (Fox, 2010) was used to calculate the variable correlation. Fig. A1.1 shows a complete depiction of the calculated correlations.



	River.Reach	Year	Meso.type	Meso.diversity	Length	Width	Depth	M.Depth	Velocity	Substrate	Embeddedness	Cover	Shadow	Wood.debries	Elevation	Slope	D.redfin	D.b.trout	D.eel	D.r.trout	D.cyprinids	Inv.density	Inv.richness	Inv.diversity	Inv.biomass
River.Reach		-6	-3	32	6	6	-31	-38	-54	3	13	-22	-28	-5	-51	-2	12	40	-30	-11	37	31	10	-21	21
Year	-6		3	-4	-18	-5	-22	-24	-46	-9	44	1	-1	-1	1	2	-26	10	-4	14	15	22	18	0	22
Meso.type	-3	3		-8	-1	0	56	66	-30	-18	21	35	7	6	-1	7	17	4	3	4	10	8	18	17	6
Meso.diversity	32	-4	-8		15	9	-6	-1	-4	-5	-11	-2	-27	-25	-24	-29	19	18	-3	-3	22	29	9	-7	25
Length	6	-18	-1	15		39	16	26	24	13	-14	0	-38	0	-36	-8	20	-1	1	-11	26	6	-7	-13	7
Width	6	-5	0	9	39		27	27	1	25	-12	-29	-56	-13	-60	-4	27	-21	-20	26	38	9	13	-3	15
Depth	-31	-22	56	-6	16	27		84	7	-7	-5	33	-3	-2	-21	19	23	-17	13	36	17	-1	14	7	1
M.Depth	-38	-24	66	-1	26	27	84		11	-2	-9	40	-1	8	-9	10	19	-15	8	22	8	4	16	17	7
Velocity	-54	-46	-30	-4	24	1	7	11		8	-47	-5	0	-7	29	-18	11	-31	22	-23	-28	-22	-19	1	-17
Substrate	3	-9	-18	-5	13	25	-7	-2	8		-22	-16	-3	-19	-4	31	-9	-2	-27	-7	-4	3	8	13	7
Embeddedness	13	44	21	-11	-14	-12	-5	-9	-47	-22		-3	-3	14	-9	-1	-3	3	4	5	11	3	6	3	5
Cover	-22	1	35	-2	0	-29	33	40	-5	-16	-3		35	9	25	4	-14	11	13	-1	-9	-3	-1	6	-10
Shadow	-28	-1	7	-27	-38	-56	-3	-1	0	-3	-3	35		19	51	29	-17	4	15	-13	-36	-17	-7	16	-17
Wood.debries	-5	-1	6	-25	0	-13	-2	8	-7	-19	14	9	19		12	-1	-5	-21	6	-2	-12	-8	3	12	-5
Elevation	-51	1	-1	-24	-36	-60	-21	-9	29	-4	-9	25	51	12		3	-34	-9	5	-38	-68	-34	-23	18	-29
Slope	-2	2	7	-29	-8	-4	19	10	-18	31	-1	4	29	-1	3		-24	4	-6	10	-26	12	10	5	8
D.redfin	12	-26	17	19	20	27	23	19	11	-9	-3	-14	-17	-5	-34	-24		2	33	-9	30	-2	-5	-15	0
D.b.trout	40	10	4	18	-1	-21	-17	-15	-31	-2	3	11	4	-21	-9	4	2		17	-25	-6	12	-9	-24	0
D.eel	-30	-4	3	-3	1	-20	13	8	22	-27	4	13	15	6	5	-6	33	17		-17	-28	-18	-16	-9	-19
D.r.trout	-11	14	4	-3	-11	26	36	22	-23	-7	5	-1	-13	-2	-38	10	-9	-25	-17		42	23	40	10	29
D.cyprinids	37	15	10	22	26	38	17	8	-28	-4	11	-9	-36	-12	-68	-26	30	-6	-28	42		28	20	-20	29
Inv.density	31	22	8	29	6	9	-1	4	-22	3	3	-3	-17	-8	-34	12	-2	12	-18	23	28		60	-14	87
Inv.richness	10	18	18	9	-7	13	14	16	-19	8	6	-1	-7	3	-23	10	-5	-9	-16	40	20	60		45	50
Inv.diversity	-21	0	17	-7	-13	-3	7	17	1	13	3	6	16	12	18	5	-15	-24	-9	10	-20	-14	45		-15
Inv.biomass	21	22	6	25	7	15	1	7	-17	7	5	-10	-17	-5	-29	8	0	0	-19	29	29	87	50	-15	

Fig. A1.1 Correlation coefficients among inputs variables. Variables codes and a short description appear in the Table 1 within the main document.

2 References

Fox, J., 2010. polycor: Polychoric and Polyserial Correlations. R package version 0.7-8.

Appendix A2

1 Selection of the starting set of variables

The step-forward approach strongly depends on the first selected variable, thus conditioning the entire process. To overcome this limitation, one MLP Ensemble was developed for each pairwise combination of variables. The best pair of variables (i.e. the one that showed the minimal Mean Squared Error, MSE) was then selected as the starting set on the step-forward procedure. Table A2.1 shows the better pairs as well as the worst for the physical habitat model (i.e. the model without biotic variables) whereas Table A2.2 shows the better pairs as well as the worst for the biotic model.

Table A2.1 Mean squared error (MSE) calculated in the selection of the starting pair of variables for the physical habitat model. Variables codes and a short description appear in the Table 1 within the main document. Bold highlight the selected pair and the control variables.

Rank	Variable 1	Variable 2	MSE
1	**Elevation**	**Embeddedness**	**0.68**
2	Embeddedness	Cover	0.68
3	Elevation	Meso.diversity	0.69
4	Elevation	Slope	0.69
5	**River.Reach**	**Year**	0.70
6	Shadow	Meso.diversity	0.70
7	Elevation	Velocity	0.71
8	**Year**	Elevation	0.71
9	**River.Reach**	Elevation	0.71
10	Depth	Elevation	0.71
.	.	.	.
.	.	.	.
.	.	.	.
113	Substrate	Wood.debris	0.85

Table A2.2 Mean squared error (MSE) calculated in the selection of the starting pair of variables for the biotic model. Variables codes and a short description appear in the Table 1 within the main document. Bold highlight the selected pair and the control variables.

Rank	Variable 1	Variable 2	MSE
1	**D.eel**	**D.cyprinids**	**0.61**
2	**Year**	D.cyprinids	0.62
3	Width	D.cyprinids	0.63
4	Elevation	D.eel	0.64
5	**River.Reach**	D.eel	0.67
6	Elevation	Embeddedness	0.68
7	Embeddedness	Cover	0.68
8	**River.Reach**	D.cyprinids	0.68
9	Elevation	Meso.diversity	0.69
10	**Year**	D.eel	0.69
.	.	.	.
.	.	.	.
.	.	.	.
265	Wood.debris	Inv.richness	0.87

2 Step-forward variable selection

To determine the variables that shape the ecological niche in a more decisive way an input variable subset was selected based on a step-forward procedure. However, instead of the usual procedure that stops when no improvement is achieved, the procedure was sustained until no more variables were available. Thereby, a MLP Ensemble for each tested combination of variables was developed. The selected variables subset and consequently, the selected MLP Ensemble was the one that presented the better performance with the lower amount of variables. To select this model the performance of the best MLP Ensemble at very iteration within the step-forward procedure and the number of variables considered at every iteration were rescaled between 0 and 1 (1 optimal), and the optimal MLP Ensemble was the one that maximized the sum of both criteria. Table A2.3 shows the selected set of variables and the corresponding criteria for the physical habitat model (i.e. the model without biotic variables) whereas Table A2.4 shows the selected set of variables, and the corresponding criteria for the biotic model. Neither correlated (r^2 < 0.5) nor collinear (*vif* < 5) combinations of variables were allowed. Consequently, as certain variables were selected some others were rejected from the step-forward procedure. Therefore, they do not appear on the corresponding table.

Table A2.3 Mean squared error (MSE) calculated for the physical habitat model in the step-forward procedure, rescaling of the calculated MSE and the number of variables. The selected model was the one that presented the maximum sum of both rescaled criteria. Variables codes and a short description appear in the Table 1 within the main document. Bold highlight the selected variables and the control variables; River.Reach and Year.

Step	Variables	MSE	Scaled MSE	Scaled Nvar	SUM	Observations
1	**Elevation + Embeddedness**	0.68	0.00	1.00	1.00	
2	**Depth**	0.43	0.41	0.89	1.30	
3	**Slope**	0.29	0.64	0.78	1.42	
4	**Cover**	**0.14**	**0.89**	**0.67**	**1.55**	**Max.**
5	Wood.debris	0.14	0.89	0.56	1.44	
6	Substrate	0.08	0.98	0.44	1.42	
7	Length	0.08	0.98	0.33	1.31	
8	**Year**	0.07	1.00	0.22	1.22	
9	Meso.diversity	0.07	1.00	0.11	1.11	
10	**River.Reach**	0.10	0.94	0.00	0.94	

Table A2.4 Mean squared error (MSE) calculated for the biotic model in the step-forward procedure, rescaling of the calculated MSE and the number of variables. The selected model was the one that presented the maximum sum of both rescaled criteria. Variables codes and a short description appear in the Table 1 within the main document. Bold highlight the selected variables and the control variables; River.Reach and Year.

Step	Variables	MSE	Scaled MSE	Scaled Nvar	SUM	Observations
1	**D.eel + D.cyprinids**	0.61	0.00	1.00	1.00	
2	**Width**	0.29	0.57	0.93	1.51	
3	**Inv.density**	0.22	0.70	0.87	1.57	
4	**Cover**	**0.13**	**0.87**	**0.80**	**1.67**	**Max.**
5	Meso.type	0.10	0.93	0.73	1.66	
6	Substrate	0.07	0.98	0.67	1.64	
7	Wood.debris	0.06	1.00	0.60	1.60	
8	Meso.diversity	0.06	1.00	0.53	1.53	
9	D.b.trout	0.07	0.99	0.47	1.46	
10	**Year**	0.07	0.98	0.40	1.38	
11	Slope	0.07	0.99	0.33	1.33	
12	Embeddedness	0.08	0.96	0.27	1.23	
13	Inv.diversity	0.08	0.97	0.20	1.17	
14	Length	0.08	0.96	0.13	1.10	
15	D.r.trout	0.10	0.94	0.07	1.00	
16	Velocity	0.09	0.95	0.00	0.95	

Appendix A3

1 Statistics of the training datasets considered within the optimal MLP Ensembles.

To increase diversity between the MLP candidates, the database was divided in 63 different training and validation datasets corresponding to all possible combinations of 66 % of the cases for training and 33 % for validation (i.e. following the *k-fold* approach). The step-forward routine was applied in the selection of the MLP candidates and thus, the optimal MLP Ensemble may not include the complete training database.

1.1 Physical habitat model

The physical habitat model selected fifteen MLPs. Consequently fifteen training datasets were involved in its development. Fig. A3.1 describes the basic statistics of every considered training dataset by the corresponding variable.

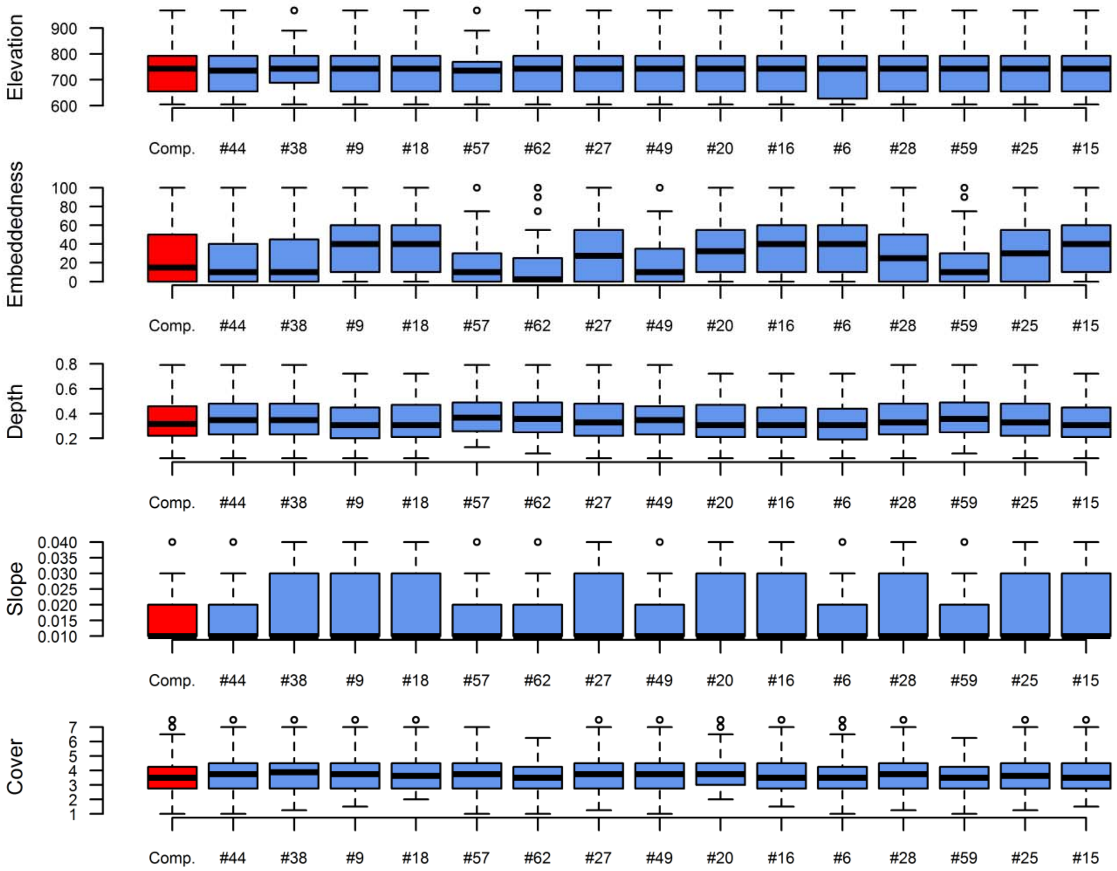


Fig. A3.1 Boxplots of the complete database (red) and the training datasets (blue) considered within the optimal physical habitat model.

1.2 Biotic model

The physical habitat model selected eight MLPs. Consequently, eight training datasets were involved in its development. Fig. A3.2 describes the basic statistics of every considered training dataset by the corresponding variable.

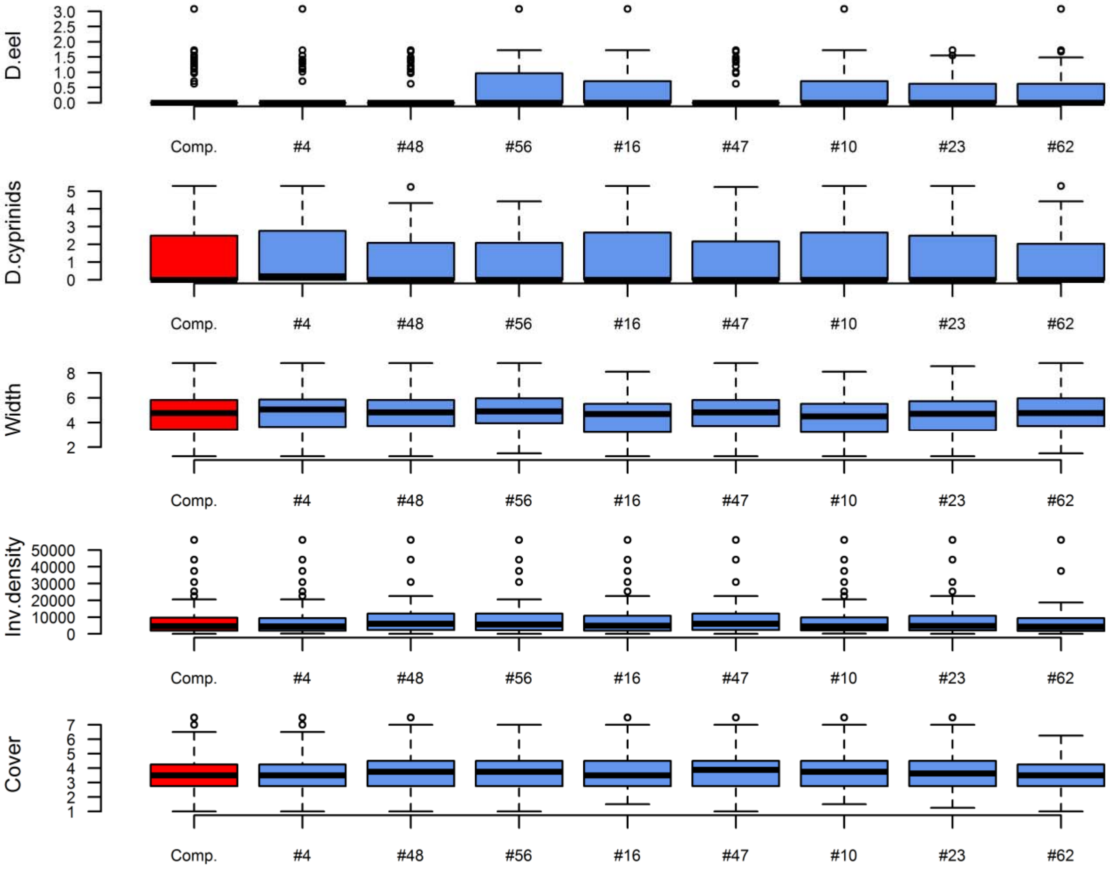


Fig. A3.2 Boxplots of the complete database (red) and the training datasets (blue) considered within the optimal biotic model.