Ecological Informatics 34 (2016) 22–34

Contents lists available at ScienceDirect

Ecological Informatics

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

Comparing four methods for decision-tree induction: A case study on the invasive Iberian gudgeon (*Gobio lozanoi*; Doadrio and Madeira, 2004)

Rafael Muñoz-Mas [a,*], Shinji Fukuda [b], Paolo Vezza [c], Francisco Martínez-Capel [a]

[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] Institute of Agriculture, Tokyo University of Agriculture and Technology, Saiwai-cho 3-5-8, Fuchu, Tokyo 183-8509, Japan
[c] International Centre for Ecohydraulics Research (ICER), University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom

ARTICLE INFO

Article history:
Received 13 January 2016
Received in revised form 18 March 2016
Accepted 24 April 2016
Available online 29 April 2016

Keywords:
Evolutionary tree
Mediterranean river
Mesohabitat suitability model
Oblique tree
R

ABSTRACT

The invasion of freshwater ecosystems is a particularly alarming phenomenon in the Iberian Peninsula. Habitat suitability modelling is a proficient approach to extract knowledge about species ecology and to guide adequate management actions. Decision-trees are an interpretable modelling technique widely used in ecology, able to handle strongly nonlinear relationships with high order interactions and diverse variable types. Decision-trees recursively split the input space into two parts maximising child node homogeneity. This recursive partitioning is typically performed with axis-parallel splits in a top-down fashion. However, recent developments of the *R* packages *oblique.tree*, which allows the development of oblique split-based decision-trees, and *evtree*, which performs globally optimal searches with evolutionary algorithms to do so, seem to outperform the standard axis-parallel top-down algorithms; CART and C5.0. To evaluate their possible use in ecology, the two new partitioning algorithms were compared with the two well-known, standard axis-parallel algorithms. The entire process was performed in *R* by simultaneously tuning the decision-tree parameters and the variables subset with a genetic algorithm and modelling the presence–absence of the Iberian gudgeon (*Gobio lozanoi*; Doadrio and Madeira, 2004), an invasive fish species that has spread across the Iberian Peninsula. The accuracy and complexity of the trees, the modelled patterns of mesohabitat selection and the variables importance were compared. None of the new *R* packages, namely *oblique.tree* and *evtree*, outperformed the C5.0 algorithm. They rendered almost the same decision-trees as the CART algorithm, although they were completely interpretable – they performed from four to eight partitions – in comparison with C5.0, which resulted in a more complex structure with 17 partitions. *Oblique.tree* proved to be affected by prevalence and it does not include the possibility of weighting the observations, which potentially discourage its actual use. Although the use of *evtree* did not suggest a major improvement compared with the remaining packages, it allowed the development of regression trees which may be informative for additional modelling tasks such as abundance estimation. Looking at the resulting decision-trees, the optimal habitats for the Iberian gudgeon were large pools in lowland river segments with depositional areas and aquatic vegetation present, which typically appeared in the form of scattered macrophytes clumps. Furthermore, Iberian gudgeon seems to avoid habitats characterised by scouring phenomena and limited vegetated cover availability. Accordingly, we can assume that river regulation and artificial impoundment would have favoured the spread of the Iberian gudgeon across the entire peninsula.

1. Introduction

The impacts of foreign fish species are recognised as a major threat to global biodiversity *via* a variety of adverse impacts, such as habitat alteration, predation, hybridisation, vectoring diseases, food web alteration and interspecific competition (Almeida and Grossman, 2012). Consequently, it is commonly known that the introduction of a foreign species in an ecosystem always poses various ecological risks (Gozlan et al., 2010). The Iberian Peninsula is considered one of the freshwater fish biodiversity hotspots in Europe (Reyjol et al., 2007), with several species at imminent risk of extinction (Leunda, 2010). Lamentably, the rate and extent of invasions in freshwater ecosystems are particularly alarming in this region, with constant reports about new successful introductions (Ilhéu et al., 2014).

From an ecological viewpoint, one species artificially moved from one basin to another in the same country could generate similar ecological outcomes (*e.g.*, increases in predation, competition or hybridisation) as a species moved across a national border. However;

Abbreviations: CART, classification and regression tree; GA, genetic algorithm; HMU, hydro-morphological unit; MSE, mean squared error; Sn, sensitivity; Sp, specificity; TSS, true skill statistic.

* Corresponding author.
E-mail address: pitifleiter@hotmail.com (R. Muñoz-Mas).

http://dx.doi.org/10.1016/j.ecoinf.2016.04.011
1574-9541/

fish species that have been introduced in other basins, within the same national borders, have benefited from a special status, for which non-native and invasive species management policies have typically not been applied (Gozlan et al., 2010). Thereby, despite the evidence that these species can impact recipient ecosystems in a similar way to foreign ones, the term translocated has been misleadingly coined to encompass such fellow species spread throughout their countries of origin (Oscoz et al., 2006; Alcaraz et al., 2014). For instance, the Iberian nase (*Pseudochondrostoma polylepis*; Steindachner, 1866) has proven to be a superior competitor and over time it has almost totally displaced the Júcar nase (*Parachondrotoma arrigonis*; Steindachner, 1866) from its historical distribution area. Another relevant example is the Iberian gudgeon (*Gobio lozanoi*; Doadrio and Madeira, 2004) (Doadrio and Madeira, 2004), which has proven to be tremendously versatile in its ecological requirements. It has been able to successfully spread across the Iberian Peninsula (Comesaña and Ayres, 2009; Ilhéu et al., 2014; Ribeiro et al., 2009) increasing the competition for the available habitat resources (Almeida and Grossman, 2012; Aparicio et al., 2013).

River systems in the Iberian Peninsula, specifically Spain, are among the most regulated in the world (García de Jalón, 1987) significantly increasing artificial impoundment and conferring water managers with an enormous capacity to manipulate the flow regime. For invasive species, risk assessment in the Iberian Peninsula has typically been addressed at the basin scale by identifying key biological traits that would facilitate successful invasions (Almeida et al., 2013) thus quantifying the degree of invasiveness of large sets of fish species (Clavero, 2011; Ribeiro et al., 2008; Ilhéu et al., 2014). However, once the invasion took place, management tools and mitigation protocols at the appropriated scale have been stressed as necessary (Gozlan et al., 2010; Sadeghi et al., 2013), otherwise unsubstantiated manipulative actions (*e.g.*, those necessary to deal with climate change-induced needs) may be favourable to non-native and invasive species.

In this regard, habitat suitability modelling based on machine learning techniques is increasingly recognised and widely applied to extract knowledge on species ecology (Fukuda and De Baets, 2012), conferring scientists and researchers with the capability to perform accurate spatial and temporal predictions (Olden et al., 2008). To date, a huge number of different techniques are available to develop these habitat suitability models – also known as species distribution models – from the relatively complex model ensembles (*e.g.*, random forests or multilayer perceptron ensembles), where several models are induced and used to perform co-ordinate predictions (*e.g.*, Fukuda et al., 2014; Muñoz-Mas et al., 2015) to the relatively simple decision-trees (*e.g.*, Leclere et al., 2011; Fukuda et al., 2014). Thus, there are several examples on habitat suitability modelling and freshwater ecology proficiently addressed with the aforementioned machine learning techniques (Leclere et al., 2011; Fukuda et al., 2014; Muñoz-Mas et al., 2015) and others, such as support vector machines (Fukuda et al., 2014; Kwon et al., 2015), fuzzy logic (Muñoz-Mas et al., 2012), as well as comparisons between them (Leclere et al., 2011; Fukuda et al., 2013; Kwon et al., 2015). However, based on such comparisons, no consensus has been reached on the optimal technique, since each modelling technique has its own unique structure and merits that may tip the balance one way or another when deciding which best fits the fundamental requirements of the problem (Lin et al., 2015).

For instance, one relevant drawback consists of the accuracy-interpretability trade-off (*i.e.* the balance between precise prediction and the capacity to easily comprehend the modelled habitat selection patterns) (Fukuda and De Baets, 2012). Accurate models, such as the ensemble ones, tend to need excessive parameterisation which makes them less interpretable (Fukuda et al., 2011b). Furthermore the internalities of every machine learning technique increase or decrease such limitations, thus those models termed as black-boxes (*e.g.*, artificial neural networks) may aggravate them (Olden and Jackson, 2002). Nowadays several of those black-box approaches allow an adequate interpretation of the modelled patterns (*e.g.*, multi-layer perceptron ensembles, Muñoz-Mas et al., 2015). Yet, remarkable differences still exist concerning the interpretability of each modelling technique. For instance, fuzzy logic-based models (Adriaenssens et al., 2004) are highly interpretable models, whereas artificial neural networks or some ensemble approaches, such as random forests (Breiman, 2001), require indirect methods to scrutinise them (Friedman, 2001).

Among these, decision-trees have been highlighted as especially suited for studies where interpretability should prevail since they typically render compact models depicted in the form of tree-like graphs (Grubinger et al., 2014). In addition, they are able to handle strongly nonlinear relationships with high order interactions and different variable types (Olden et al., 2008; Grubinger et al., 2014). Furthermore, its induction (training) has demonstrated to be significantly faster than other machine learning techniques (*e.g.*, artificial neural networks) (Olden et al., 2008), which is especially appealing for big data. Consequently, there are several benchmarking studies that used decision-trees to model the habitat preferences of alpine fish (Vezza et al., 2014) and invasive non-native species (*e.g.*, Sharma et al., 2009).

The very basic principle in decision-tree induction consists of splitting the training dataset using recursive partitioning algorithms, by which the data set is iteratively divided into two parts maximising homogeneity (*e.g.*, minimising an impurity measure) in the child nodes (Grubinger et al., 2014). This splitting or partitioning typically starts from the largest discriminant split to the least one and it is applied in a hierarchical fashion to each of the new branches of the tree until the maximum number of allowed partitions or any other constrain is achieved (Vezza et al., 2015).

Several methods for decision-tree induction exist, from the old fashion CHAID (Chi-Square Automatic Interaction Detector) (Kass, 1980), which is restricted to categorical variables, to novel methods that use memetic algorithms to induce globally optimal oblique trees (Czajkowski and Kretowski, 2013). Each impurity measure has its own merits and demerits, which define different optimisation problems (Cantú-Paz and Kamath, 2003). Among possible induction methods, one of the most popular algorithms for freshwater fish studies (Fukuda et al., 2014; Parasiewicz et al., 2012), is classification and regression trees (CARTs) (Breiman et al., 1984), which later triggered the development of other tree-based ensemble machine learning techniques, such as random forests (Breiman, 2001). Another series of popular algorithms for decision-tree induction are those conformed of the Iterative Dichotomiser (known as ID3) and the superseding C4.5 and C5.0 algorithms developed by Quinlan (1992) with examples on fish and freshwater ecosystems (*e.g.*, Baxter and Shortis, 2002; D'heygere et al., 2006). However all of the aforementioned algorithms separate the feature space by axis-parallel hyperplanes, which may be sub-optimal (Truong, 2009) and ecologically unreliable because they render stair-like decision surfaces (Menze et al., 2011).

Oblique splits may overcome these limitations, producing interpretable and more accurate trees with decision boundaries less biased by geometrical constraints of the base learner (Murthy et al., 1994; Truong, 2009). However, in oblique tree induction the number of possible splits grows extremely quickly with sample size and number of variables (Truong, 2009). Finding the best oblique tree is a NP-complete problem (Heath et al., 1993) and consequently the oblique inducers require greater computational resources. As a consequence oblique tree inducers use heuristics to find proficient partitions since exhaustive searches are unaffordable (Cantú-Paz and Kamath, 2003). There are several approaches to develop oblique trees from the very simple Breiman's perturbation approach (Breiman et al., 1984), to others based on logistic regression (Truong, 2009) or simulated annealing (Heath et al., 1993). Despite oblique trees represent a major improvement over axis-parallel ones; they typically perform the partition of the input space also in a top-down fashion, without consideration of nodes further down the tree (*e.g.*, Breiman et al., 1984; Murthy et al., 1994). Sequentially induced trees can be far from the optimal solution thus global searches using evolutionary strategies can lead to much

more compact and accurate decision-trees (Grubinger et al., 2014; Cantú-Paz and Kamath, 2003). Unfortunately, such oblique and evolutionary approaches are rarely used, principally because they have not been accessible for potential users (Truong, 2009; Grubinger et al., 2014).

To date, several axis-parallel approaches are available in *R* software and accordingly CART and C4.5/C50 are actually implemented in several packages such as *tree* (Ripley, 2015), *rpart* (Therneau et al., 2015) or *C50* (Kuhn et al., 2015). The ecosystem of user-contributed *R* packages has been growing steadily at a significantly fast rate (German et al., 2013). Therefore, two new packages for oblique tree (*oblique.tree*, Truong, 2013) and evolutionary tree (*evtree*, Grubinger et al., 2014) induction are already available.

The present study compared such novel packages, *oblique.tree* and *evtree*, with two well-known axis-parallel top-down approaches, CART and C5.0, which were developed with the *tree* and the *C50* packages respectively. The comparison was carried out modelling the presence-absence (suitability) of the invasive Iberian gudgeon (*G. lozanoi*) at the mesohabitat scale. The accuracy, the modelled patterns of habitat selection, the tree complexity and the variable importance were compared between the four decision-tree induction techniques. Fish ecology and the implication of future management actions were briefly discussed.

2. Methods

2.1. Iberian gudgeon ecology

Formerly the distribution area of the Iberian gudgeon encompassed the Ebro and Bidasoa River Basins (Doadrio, 2002), thus the remaining populations in the Iberian Peninsula should be considered non-native. Its spread has been caused by introduction as fishing bait, and due to inter-basin water transfers (Clavero and García-Berthou, 2006). Despite its successful dispersion, studies describing its habitat preferences are still scarce (Lobon-Cervia et al., 1991; Miñano et al., 2003; Ilhéu et al., 2014). This small schooling cyprinid – maximum body length ≈ 150 mm (Doadrio and Madeira, 2004) – feeds largely on macroinvertebrates (Oscoz et al., 2006) and inhabits stretches of intermediate elevation and flow velocity, preferably with sandy substrates (Doadrio, 2002). It has been reported to colonise lentic environment such as reservoirs (Miñano et al., 2003) and the species is sensitive to significant reductions in water quality caused by pollution (Ilhéu et al., 2014).

2.2. Habitat data

Input data were retrieved from previous studies performed in four unregulated segments (*i.e.*, upstream of the Contreras dam) of the Cabriel River (eastern Iberian Peninsula) from 2006 to 2009 (Costa et al., 2012; Vezza et al., 2015) (Fig. 1). The Cabriel River basin is a low-density populated area that has been affected by a marked human depopulation during the last fifty years (Instituto Nacional de Estadística, 2013). In accordance with this, the surveyed area has been categorised as pristine – very low pressure level – by the water administration (Confederación Hidrográfica del Júcar, 2005). Accordingly, impacts or differences regarding the physical–chemical predictors throughout the study area were considered negligible and the study focused only on the physical habitat. Furthermore, the Iberian gudgeon only cohabits with another invasive species (the Iberian nase) at C3 and C4 (Alcaraz et al., 2014; Vezza et al., 2015). Therefore the uneven distribution of the species lead to preferable modelling the presence–absence of the target species instead of using multivariate regression trees to predict multiple species distributions (*e.g.* Wilkes et al., 2015).

The presence–absence of the Iberian gudgeon and the characteristics of the physical habitat were surveyed at the mesohabitat scale, equalling mesohabitats with hydro-morphological units (HMUs). In particular, each year, the four river segments were stratified in five different

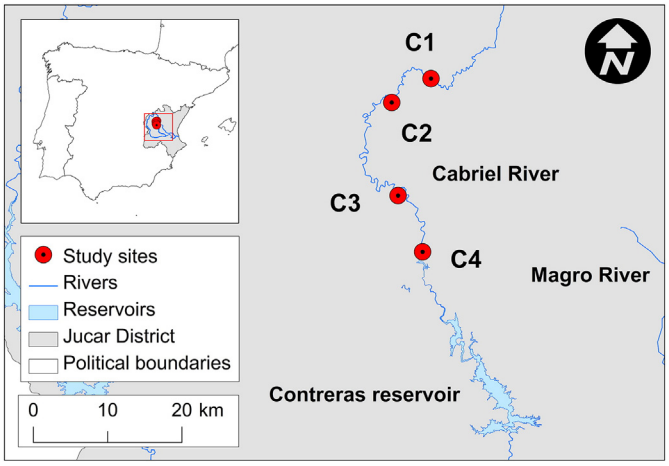

Fig. 1. Location of the study sites within the upper part of the Cabriel River (Jucar River basin District — eastern Iberian Peninsula).

types of HMUs, namely: pool, glide, run, riffle, and rapid (Costa et al., 2012; Vezza et al., 2015) and the sequences of HMUs were selected to encompass complete HMUs summing river lengths that slightly exceeded 1 km. After the three campaigns, the percentages of occurrence of each HMU class, which were used as an ordinal input variable (HMU type), were of 32%, 5%, 2%, 47%, and 14% respectively. Once the set of HMUs was selected, three groups of dimension-related, flow-related and cover-related attributes were measured.

The dimensions of the HMUs were measured in terms of length, measured with CMII Hip Chain (CSP Forestry Ltd. Alford, Scotland), average width, measured with laser distance metre DISTO A5 (Leica Geosystems, Heerbrugg, Switzerland) obtained from four to eight cross-sections, mean depth (depth), measured with a wading rod and calculated from 20 to 40 point measurements at a rate of 5 measurements per transect, maximum depth (max. depth) measured in the corresponding point and, in case of significant discontinuities between HMUs, the offset depth was also measured with the wading rod (Vezza et al., 2015). Then the areas and volumes of the HMUs were calculated by considering the corresponding length, width and depth.

The flow rate at the time of the survey was gauged and was used to calculate the mean flow velocity (velocity). The HMU gradient was measured with a Haglöf HEC Electronic Clinometer (Haglöf Sweden AB, Långsele, Sweden) and the backwater and pocket water areas were recorded if present; *i.e.* if waters were visibly stagnated or backed up by lateral (backwaters) or big rocks (pocket waters) obstructions. The percentages of substrate types were visually estimated following a simplified classification from the American Geophysical Union (Muñoz-Mas et al., 2012) and summarised in the substrate index (Mouton et al., 2011), which ranges from 0 (vegetated silt) to 8 (bedrock). The percentage of the HMU area covered by silt and mud was also recorded (% embeddedness).

The cover-related group of attributes was visually estimated. It included % shade (percentage of the overall HMU's area), % undercut banks (percentage of the HMU's length), % aquatic vegetation (percentage of the overall HMU's area) and % reeds (percentage of the HMU's length). Additionally the cover index ranging from 0 to 10 (García de Jalón and Schmidt, 1995) was determined to characterise the available refuge due to caves, shade, large substrate, aquatic vegetation and water depth. The amount of big boulders (# boulders) and woody debris (# woody debris) was counted and considered as input variables. The density of these countable items was calculated by dividing their number by the HMU area (*i.e.* density of woody debris and density of boulders). Finally, for pocket waters and backwaters the same procedure was followed (‰ pocket waters and % back waters) (Table 1).

The biological survey was concomitantly performed with the physical habitat survey by snorkelling the aforementioned HMUs. Two divers

Table 1
Summary and units of input variables collected in the upper segment of the Cabriel River.

Variable	Type	Min.	1st qu.	Median	Mean	3rd qu.	Max.	Units
Length	Continuous	5.3	22.95	37.7	47.84	64.9	191.6	m
Width	Continuous	2.7	6.53	7.98	8.509	10.36	20.03	m
Area	Continuous	25.92	154.5	317.9	732	694.4	11970	m^2
Volume	Continuous	10.98	115.4	253.3	748.5	668.4	12860	m^3
Depth	Continuous	0.27	0.598	0.85	0.937	1.17	3.52	m
Max. depth	Continuous	0.34	1	1.39	1.498	1.892	4	m
Offset depth	Continuous	0.1	0.48	0.66	0.76	0.95	3.52	m
Velocity	Continuous	0.04	0.13	0.225	0.275	0.36	1.05	m/s
Gradient	Continuous	0	0.01	0.02	0.022	0.03	0.093	m/m
% aquatic vegetation	Discrete	0	10	15	21.47	25	95	%
Substrate index	Continuous	0.3	3.65	4.4	4.235	5.1	8	–
Density of woody debris	Continuous	0	0	0	0.003	0	0.04	$\#/m^2$
# woody debris	Discrete	0	0	0.25	0.854	1	9	#
% embeddedness	Discrete	0	0.05	0.125	0.206	0.35	1	%
% shade	Discrete	0	0.2	0.35	0.392	0.6	1	%
Cover index	Continuous	2.5	4.25	5	5.41	6.25	9	–
Pocket waters	Continuous	0	0	0	0.035	0.01	1.3	m^2
‰ pocket waters	Discrete	0	0	0	0.203	0.012	5.23	‰
% backwaters	Discrete	0	0	0.01	0.04	0.05	0.53	%
Back waters	Continuous	0	0	3.8	12.54	15	389.1	m^2
Density of boulders	Continuous	0	0	0.01	0.028	0.04	0.28	$\#/m^2$
# boulders	Discrete	0	0	3	6.153	9	87	#
% undercut banks	Discrete	0	0	0.05	0.174	0.25	1	%
% reeds	Discrete	0	0.13	0.3	0.341	0.5	0.98	%

conducted the underwater counts in three independent passes from downstream to upstream (Costa et al., 2012; Vezza et al., 2015). Divers were trained to maintain the fish sampling effort constant during the three replicate counts, ensuring a reasonably uniform probability of detection (Schill and Griffith, 1984). In order to keep each pass independent (*i.e.* unaffected by previous passes) a time delay of about 2 h was programmed between replicate counts (Bain et al., 1985). The snorkelling technique was chosen due to the hydraulic and morphological characteristics of the river (*i.e.* clear water and deep pools, max. depth *ca.* 4 m) and the accuracy as previously demonstrated (Heggenes et al., 1990).

Presence–absence modelling was the selected choice, as it is likely to yield robust results and allows the potential influence of density-dependent phenomena to be ruled out (Fukuda et al., 2011a). In addition it better fits the probabilistic-like outputs required in physical habitat simulation studies (Bovee et al., 1998). Because the spatial dependency in fish distributions was evaluated to be random with no evidence of spatial autocorrelation (Vezza et al., 2015), the data were pooled with no major consideration about year, study site or HMU sequence. In the end Iberian gudgeon was observed in 100 out of 268 HMUs resulting in high prevalence (*i.e.* ratio of presence cases over the entire dataset) of 0.37.

2.3. Decision-tree induction

The parameters controlling complexity (Olden et al., 2008) and the number of input variables (D'heygere et al., 2006) determine the capability of the decision-tree to accurately predict unseen data (*i.e.* generalisation). Complexity can be restricted *a priori* or *a posteriori* (Olden et al., 2008). For instance, *a priori* the maximum number of terminal nodes can be restricted (Olden et al., 2008) whereas pruning has become the most popular approach *a posteriori* because knowing beforehand when to stop tree growing is problem-dependent (Galathiya et al., 2012). However, there are several pruning approaches, which may lead to dissimilar decision-trees (Esposito et al., 1997). Consequently, the use of *a priori* constraints gained certain relevance (Garofalakis et al., 2000). The group of selected input variables also control complexity because large sets tend to increase the number of partitions and terminal nodes (D'heygere et al., 2006; Inza et al., 2004). Several approaches exist to address the variable selection problem such as filters (Inza et al., 2004), greedy approaches (Sadeghi et al., 2013) and wrappers (Inza et al., 2004). Filters evaluate inputs prior to modelling by scrutinising input–output relationships and thus are irrespective to the modelling approach eventually used (Inza et al., 2004). Conversely greedy or wrapper approaches select the variables on the basis of their predictive capability through the selected modelling approach (Sadeghi et al., 2013; Inza et al., 2004). Greedy approaches rely on iteratively adding or removing inputs whereas wrappers do so by globally searching optimal subsets (D'heygere et al., 2006). Genetic algorithms (GAs) (Holland, 1992), which are based on the process of natural selection – selection, reproduction and mutation – (Huang and Wang, 2006), have demonstrated a proficient global searching strategy of the best set of inputs for decision-tree induction (D'heygere et al., 2006; Inza et al., 2004; Sadeghi et al., 2013) and they are suited to perform searches over complex parameter structures (Olden et al., 2008). Consequently GAs have demonstrated to be proficient wrappers simultaneously searching optimal parameters and input variable sets (Huang and Wang, 2006).

The optimisation of the decision-tree parameters and the best inputs subset were performed coordinately with the GA comprised in the *rgenoud* package (Mebane and Sekhon, 2011) (Fig. 2-A). GAs encode each combination of parameters and variables in sequences (chromosomes) where each value corresponds to a gene whereas the group of chromosomes is the population. During every iteration of the GA, one decision-tree is developed for each chromosome and the evolution of the population takes place in accordance with the accuracy related to every chromosome, which conditions its probability of selection (*i.e.* the probability to be directly transferred from generation to generation) and of reproduction (*i.e.* recombination with other competing chromosomes). Finally, to adequately sampling the searching space, a relevant proportion of chromosomes are also mutated (*i.e.* their values are randomly modified). The evolution halts if the maximum number of generations is achieved or if the GA is unable to find a better solution after a specified number of generations. The chromosomes were composed of integers; the first part, which varied in length, encompassed the decision-tree parameters (covering adequate ranges) whereas the second part was composed of a bit string of length equalling the number of variables within the training dataset (Fig. 2-B). The parameters that required real numbers (*e.g.* the *alpha* parameter) were obtained by dividing the corresponding gene by 100 (*i.e.* accuracy = 0.01).

The *rgenoud* function presents 9 different operators driving the optimisation (Mebane and Sekhon, 2011). To avoid premature convergence

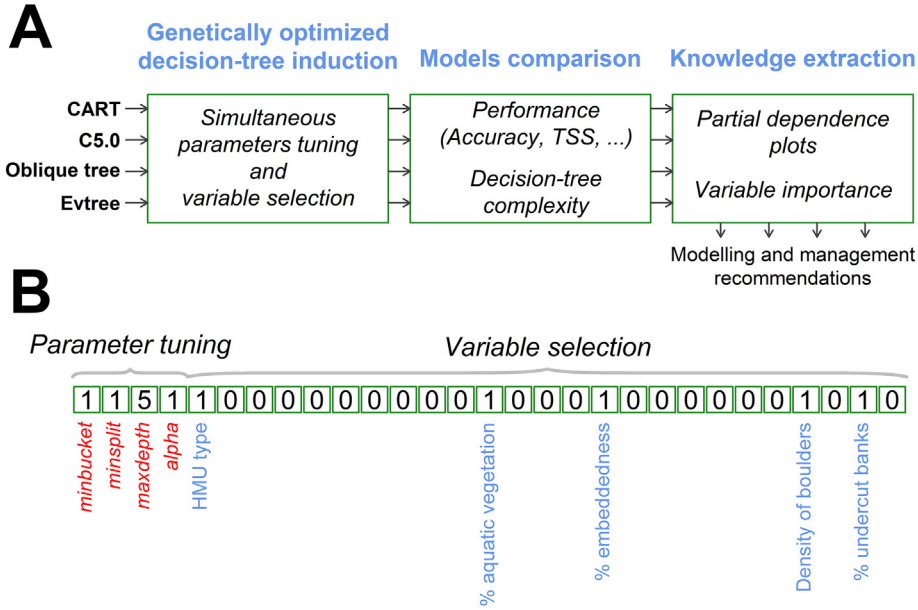

Fig. 2. Flowchart (A) and example of the structure of the chromosomes used in the genetic optimisation (B). The depicted chromosome (B) shows one of the tested alternatives for the *evtree* approach for decision-tree induction. The parameter tuning part set the minimum number of cases in a terminal node (*minbucket*) and to allow the node split (*minsplit*) to 1, the maximum number of splits (*maxdepth*) to 5 and the alpha parameter to 0.01 (*i.e.* 1/100). The selected variables in the next part were those encoded with a 1 (they appear labelled below the chromosome) whereas those encoded with a 0 were ruled out (for clarity, they are not labelled in the figure).

on sub-optimal solutions (Fogel, 1994) it is necessary to balance the population diversity and the selection pressure (Pandey et al., 2014). Consequently, the cloning operator that controls selection was restricted (0.25) whereas those operators addressed to increase diversity (*i.e.* those that control mutation and crossover) were favoured with relatively high values (0.75, 0.75 and 0.35) (see Table 2 for a complete depiction of parameters' settings). However, the operators leading genes towards the extreme values (boundary mutation and whole and standard non-uniform mutation) were kept low since (as most of the chromosome was composed of bits) they only switch their value to the opposite option. Finally, the population size and the number of generations were set at 500 and the optimisation halted after 50 generations without improvement.

The objective function corresponded to the maximisation of the true skill statistic (TSS) following a repeated *k*-fold scheme (Borra and Di Ciaccio, 2010). However, owing to the number of decision-trees potentially tested, instead of the standard fivefold or tenfold cross-validation, which may also be sub-optimal (Arlot and Celisse, 2010), a three times threefold cross-validation ($3\times3_{cross-validation}$) scheme was followed because it demonstrated to be adequate to induce genetically optimised decision-trees (Stein et al., 2005) (Fig. 2-A). Every fold presented a prevalence similar to the original dataset whereas the number of variables was restricted by favouring its use in every of the nine developed decision-trees (*i.e.* ($\overline{TSS\times\#\ variables\ used/\#\ variables\ selected}$). Unlike previous studies (Sadeghi et al., 2013) and to avoid redundancy in the input data, correlated (($r^2 > 50\ \%$)) combinations of variables were not allowed. 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 variable correlations (Appendix A in the online version at http://dx.doi.org/10.1016/j.ecoinf.2016.04.011.). Following previous studies (*i.e.* Fukuda et al., 2013), once the optimal parameters and variables were obtained, a single decision-tree was calculated to inspect its predictive capability and differences with the cross-validated decision-trees.

Table 2
Genetic algorithm (*rgenoud*) parameter settings.

Number	Operator	Setting
1	Cloning	0.25
2	Uniform mutation	0.75
3	Boundary mutation	0.15
4	Non-uniform mutation	0.10
5	Polytope crossover	0.15
6	Simple crossover	0.75
7	Whole non-uniform mutation	0.00
8	Heuristic crossover	0.35
9	Local-minimum crossover	0.00

2.3.1. Classification and regression tree — CART

The CARTs were developed with the package *tree* (Ripley, 2015). In addition to the input variables subset four parameters were optimised: *mincut*, *minsize*, *mindev* and *split.impurity*, which respectively correspond to the minimum number of observations to include in either child nodes, the smallest allowed node size, the minimum ratio between the within-node and the root node for the node to be split (default set to 0.01), and the impurity index used (*i.e.* deviance or Gini) with larger values of *mincut*, *minsize* or smaller of *mindev* reducing the risk of over-fitting. The tested values ranged from 1 to 236 (*i.e.* the maximum number of training data in a given fold) for *mincut* and *minsize*, from 1 to 1000 for *mindev*, which was pertinently divided by 1000 instead of 100 (effective range between 0.001 and 1), and 0 or 1 for *split* (*i.e.* deviance or Gini) (Table 3).

2.3.2. Quinlan's decision-trees — C5.0

Quinlan's algorithms render axis-parallel decision-trees similar to CARTs, although they differ in the approach used to determine the ultimate splits (Quinlan, 1992). Thereby, C4.5/C5.0 use the normalised information gain (difference in entropy) to select the optimal splits (Quinlan, 1992) and, unlike CARTs, the ultimate partition is determined anytime in a forward/backward procedure. First, the tree is completely grown until the *a priori* constraints are met and then those branches of minor importance are pruned in accordance with the confidence factor. Further, C4.5/C5.0 introduced an alternative compact structure of the former tree consisting of a list of rules of the form IF-THEN sequences, where rules for each class are grouped together — if a case fulfils a rule, it is assigned to the corresponding category otherwise it is assigned to the default class (Quinlan, 1992). The package *C50*

Table 3
Summary of the optimised parameters, with the considered ranges, for the four decision-tree induction methods.

Method	Parameter	Bounds		Description
		Lower	Upper	
CART	*mincut*	1	236	Minimum number of observations to include in either child node
	minsize	1	236	Smallest allowed number of observation in a terminal node
	mindev	0.001	1	Minimum within-node and root node ratio of deviation for the node to be split
	split	deviance	Gini	Splitting criterion to use
C5.0	*rules*	Yes	No	Should the tree be decomposed into a rule-based model?
	subset	Yes	No	Should the model evaluate groups of discrete predictors for splits?
	bands	0	999	Group the rules into the specified number of bands
	noGlobalPruning	Yes	No	Should the global pruning step to simplify the tree to be toggle?
	CF (Confidence Factor)	0	1	Threshold of allowed error in data while pruning the decision tree
	minCases	1	236	Smallest allowed number of observation in a terminal node
	fuzzyThreshold	Yes	No	Should C5.0 evaluate possible advanced splits of the data?
Oblique tree	*mincut*	1	236	Minimum number of observations to include in either child node
	minsize	1	236	Smallest allowed number of observation in a terminal node
	mindev	0.001	1	Minimum within-node and root node ratio of deviation for the node to be split
	split.impurity	deviance	Gini	Splitting criterion to use
Evtree	*minsplit*	1	236	Minimum number of observations to include in either child node
	minbucket	1	236	Smallest allowed number of observation in a terminal node
	maxdepth	0	15	Maximum number of nodes in the tree
	alpha	0	10	Factor regulating regulates the complexity tree size
	ntrees	25	625	Population size ($25 \times$ *# variables selected*)
	niterations	100	2500	Number of generations run before premature stopping ($100 \times$ *# variables selected*)
	psplit,	0.2		Evolutionary splitting operator for terminal nodes
	pprune,	0.2		Evolutionary pruning operator of internal nodes
	pmutatemajor,	0.2		Evolutionary mutant operator of variables and values for internal nodes
	pmutateminor	0.2		Evolutionary mutant operator of values for internal nodes
	pcrossover	0.2		Evolutionary mutant operator to exchange branches

(Kuhn et al., 2015) was used to develop a single decision-tree based on Quinlan's C5.0 (1992). Although C5.0 makes several improvements in regard to C4.5, especially the option of development of ensembles of decision-trees by recursively boosting the subsequent trees (*i.e.* the subsequent trees are trained more intensely on the data that presented a greater misclassification rate), we restricted the developed model to one single decision-tree to maintain the interpretability because the ensemble typically behave as a black box. Seven parameters were optimised in addition to the input variables subset; *rules*, *subset*, *bands*, *noGlobalPruning*, *CF* (confidence factor), *minCases* and *fuzzyThreshold*. The parameter *rules* determines if C5.0 should convert the decision-tree into IF-THEN rules whereas if doing so *bands* controls the ultimate amount of developed rules with lower values favouring generalisation. *Subset* controls whether groups of discrete predictors for splits should be evaluated. *MinCases* controls tree growth (*i.e.* the smallest allowed terminal node size) and *CF* controls the intensity of the ultimate pruning whereas *noGlobalPruning* disables the latter step thus large values of *MinCases* and low values of *CF* favour generalisation whereas enabling *noGlobalPruning* has the opposite effect. Finally, *fuzzyThreshold* controls the way predictions are done by dividing each split into three ranges and if a given case lies in the middle range, two of the three branches are investigated and the results combined probabilistically. Finally, C5.0 can automatically winnow the input variables to remove those that are irrelevant; however, this option was disabled because the optimal variables subset was sought by means of the GA. The tested values ranged from 0 to 999 for *bands*, from 0 to 100 for *CF* and from 1 to 236 for *minCases* whereas the remainder parameters ranged from 0 to 1 (*i.e.* false or true) (Table 3).

2.3.3. Oblique tree

The oblique trees were developed with Truong's *oblique.tree* package (Truong, 2013), which is able to develop decision-trees mixing oblique and axis-parallel splits This approach infers the oblique splits by developing linear decision boundaries (linear classifiers) with logistic regression. Therefore, for every split the *oblique.tree* function tests $2^{R-1} - 1$ logistic regression models (where R is the number of residual classes) and selects the one that maximises the separation between the two classes (Truong, 2009). As with the other approaches, the process continues until leaves are homogeneous or the maximum number of allowed splits is achieved. *Oblique.tree* presents exactly the same parameters as the CART algorithm implemented in the package *tree* (Ripley, 2015). Therefore they were set as in the previous case, *mincut* and *minsize* ranged from 1 to 236 and *mindev* from 1 to 1000; whereas two values were allowed for *split.impurity*, which corresponded to deviance or Gini (Table 3). The variable selection was performed simultaneously with the parameter optimisation then, the use of axis-parallel and oblique splits was enabled (*oblique.splits* = "*on*") but the methods to perform variable selection (*e.g.* based on AIC or BIC) were disabled because the optimal variables subset was sought by the GA.

2.3.4. Evolutionary tree — evtree

The *evtree* function in the package *evtree* (Grubinger et al., 2014) follows the same principles described for the GA, although the genes – instead of integers representing parameters and variables – encode the splits (variable and value) of the decision-tree. Therefore *evtree* presents two groups of parameters; the first parameterises the evolutionary process and is used internally by the function and the second constrains the decision-tree eventually rendered. Unlike *rgenoud*, the *evtree* evolution is controlled by five parameters, *psplit*, *pprune*, *pmutatemajor*, *pmutateminor* and *pcrossover*, in addition to the population size and the number of generations without improvement that are run before stopping. If *psplit* is applied one random terminal-node is selected and an alternative split is reassigned, *pprune* chooses a random internal node and prunes it into a terminal node, *pmutatemajor* selects a random internal node and changes the split variable and value whereas *pmutateminor* changes the split value but not the splitting variable. Finally, *pcrossover* exchanges branches between two trees. We assumed that the population and the number of iterations run before stopping should be larger in those trials that involved a larger amount of input variables thereby the population was set to $25 \times$ *# variables selected* and the number of iterations varied following $100 \times$ *# variables selected*. Finally, the parameters driving the evolution were all set to 0.2 because all of them hinder premature convergence. Regarding the constrains applied to the decision-tree finally rendered, *evtree* presents four parameters; *minbucket*, *minsplit*, *maxdepth* and *alpha*, which control the minimum number of cases in a terminal node (*i.e.* leaf), the minimum

number of cases in a branch to split it, the maximum number of splits and the ultimate complexity of the tree. In accordance with the premises described above *minbucket* and *minsplit* ranged from 1 to 236, *maxdepth* from 1 to 15 and *alpha*, which was pertinently divided by 100, from 1 to 1000 (Table 3).

2.4. Models comparison

Model transparency is fundamental to rule out ecologically unreliable models (Austin, 2007). Accordingly, on an equal footing (*e.g.* similar accuracy), simple models are preferred over complex ones because they allow better interpretation (Grubinger et al., 2014; Truong, 2009; Wu et al., 2008). To date several approaches exist to quantify decision-tree complexity (*e.g.*, Breiman et al., 1984; Murthy et al., 1994) but each one is addressed to the corresponding approach. To allow comparison, tree complexity was evaluated following Truong's (2009, 2013) approach, which consists of aggregating the number of variables involved in every split plus one (if only axis-parallel splits are used, this approach coincides with the number of leaves) (Fig. 2-A). This approach allowed comparison with *C5.0* regardless of the ultimate nature of the rendered model, tree-like or rule-based, because the path to each leaf can be assimilated to one rule. In that case, the number of splits in each rule plus the default value was also calculated because we realised that although *C5.0* is able to render compact rule sets, these rules can be highly complex.

In accordance with the two types of decision-trees potentially obtained, tree-like or rule-based, the tree-like structure was not scrutinised (they can be consulted in the Appendix B in the online version at http://dx.doi.org/10.1016/j.ecoinf.2016.04.011.) and the relationship between the input variables and the probability of the presence was graphically characterised with the partial dependence plots (Friedman, 2001), as implemented in the package *randomForests* (Liaw and Wiener, 2002) (Fig. 2-A). Partial dependence plots depict the average of the response variable *vs.* a predictor variable and account for the effects of the remaining variables within the model (Friedman, 2001). Consequently, partial dependence plots are a useful way to visualise the marginal effect of the selected variables on the predicted probability of presence (Cutler et al., 2007). The function being plotted is defined as:

$$f\sim(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}) \quad (1)$$

where n corresponds to the amount of points in which the function is being plotted, x is the variable for which partial dependence is sought, and x_{iC} are the remaining variables in the dataset. In this case f corresponded to the predictions exerted by the decision trees and the partial dependence was computed for each of 50 equally spaced points over the range of each examined variable.

Furthermore, the relative importance of input variables is useful in guiding conservation policy, monitoring and sampling strategies, and formulation of testable scientific hypotheses (Kemp et al., 2007), however, there is not a unified approach for every of the four tested approaches. Therefore, we evaluated the variable importance following Kemp's et al. (2007) approach, which is based on the perturbation of the inputs and thus is irrespective of the model structure (Fig. 2-A). It consists of (i) sequentially feeding the model with the test set but replacing the values of the target input by uniformly distributed random values in the interval (0.1, 0.9), the range over which the model was originally trained, (ii) calculating each time the performance criteria (TSS), and (iii) repeating the procedure for each input parameter. The calculated TSS is then compared with the reference value (*i.e.* the TSS obtained in the corresponding fold) and the variables importance is scrutinised by developing box-plots of the ΔTSS. The most important variable shall produce the largest ΔTSS whereas the least important the smallest.

3. Results

3.1. Performance

Based on the optimum parameters sought with the GA depicted in the Table 4, the four techniques presented practically the same values of the performance criteria during the cross-validation phase but C5.0 that presented marginally superior performance (Table 5). All the techniques proved underpredictive (sensitivity < specificity), especially C5.0, and such phenomenon increased for the models developed with the complete dataset (overall). In that case C5.0 presented significantly better performance than the remaining techniques. The lapse of the optimisation of CART and C5.0 was of few minutes whereas for the oblique tree it increased one order of magnitude and for the evtree, by two orders. For the oblique tree the number of tested individuals was *ca.* 30% larger than for the remaining techniques. Although in accordance with the lapse to optimise one single oblique tree, this was not considered the main reason for such differences in the lapses in the optimisation.

3.2. Decision-tree complexity

CART and evtree rendered the simplest decision-trees with only 4 leaves whereas oblique tree presented a marginally larger complexity because it presented few oblique splits (one to three) (Fig. 3). The optimal C5.0 consisted of a set of rules (*ca.* 13 rules plus the default value) nevertheless it presented the largest complexity because these rules included large sets of conditions, as a consequence the complexity rose up to *ca.* 40 different splits (see Appendix B in the online version at http://dx.doi.org/10.1016/j.ecoinf.2016.04.011. for a tree-like depiction of the models).

3.3. Variable selection and partial dependence plots

The optimal CART selected two variables % of aquatic vegetation and density of boulders (Fig. 4). The Iberian gudgeon selected preferably HMUs with small % of aquatic vegetation and density of boulder although it avoided the HMUs without aquatic vegetation.

In the optimal C5.0, seven predictors, namely HMU type, width, % of aquatic vegetation, % of embeddedness, % of backwaters, density of boulders and % of back waters were selected (Fig. 5). The Iberian gudgeon preferably selected pool-type HMUs of large to intermediate width and small % of aquatic vegetation. The fish had a preference for straight segments (low % of backwaters) with depositional area (high % of embeddedness) and low density of boulders. The Iberian gudgeon slightly preferred non-incised HMU (low % of undercut banks).

Table 4
Optimum parameter values for the four decision tree algorithms.

Method	Parameter	Optimum value
CART	*mincut*	7
	minsize	18
	mindev	0.039
	split	deviance
C5.0	*rules*	Yes
	subset	No
	bands	56
	noGlobalPruning	No
	CF (Confidence Factor)	0.44
	minCases	2
	fuzzyThreshold	No
Oblique tree	*mincut*	4
	minsize	19
	mindev	0.234
	split.impurity	Gini
Evtree	*minsplit*	22
	minbucket	68
	maxdepth	6
	alpha	1.71

Table 5
Model performance for the four selected decision tree algorithms. Summary of the; true skill statistic (TSS), sensitivity (Sn), specificity (Sp) and mean squared error (MSE) calculated during $3\times3_{cross-validation}$ (nine models) and for the ultimate models, lapse of the optimisation and number of tested combinations of parameters and variables (chromosomes).

	Cross-validation				Overall				Time [min]	# tested individuals
	TSS	*Sn*	*Sp*	*MSE*	TSS	Sn	Sp	MSE		
CART	0.65 ± 0.18	0.82 ± 0.14	0.84 ± 0.08	0.15 ± 0.05	0.61	0.78	0.83	0.15	2.54	8983
C5.0	0.69 ± 0.14	0.79 ± 0.12	0.89 ± 0.06	0.15 ± 0.06	0.83	0.86	0.97	0.08	5.35	8230
Oblique tree	0.65 ± 0.18	0.82 ± 0.13	0.83 ± 0.09	0.15 ± 0.05	0.62	0.80	0.81	0.14	42.32	10901
Evtree	0.65 ± 0.18	0.82 ± 0.14	0.84 ± 0.08	0.14 ± 0.04	0.61	0.78	0.83	0.15	332.47	8208

The optimal oblique tree also selected % of aquatic vegetation and density of boulders. While smoother partial dependence plots were obtained from the oblique tree, practically the same pattern of habitat selection was observed in the partial dependence plots for CART; Iberian gudgeon selected HMUs with small % of aquatic vegetation and density of boulders (Fig. 6). The 3D plot highlighted the larger flexibility to adjust the discriminant hyperplane.

The optimal evolutionary tree (evtree) selected the same variables as the CART and the oblique tree (Fig. 7). Moreover the modelled preferences were similar; the Iberian gudgeon selected the HMUs with small % of aquatic vegetation and low density of boulders and avoided the HMUs without aquatic vegetation.

3.4. Variable importance

The four techniques concomitantly indicated % of aquatic vegetation as the most important variable followed by density of boulders (Fig. 8). C5.0 selected five additional variables of minor importance, which, in decreasing order, presented the following rank, HMU type, % undercut banks, width, % backwaters and % embeddedness.

4. Discussion

4.1. Comparison of the decision-tree induction methods

Four different *R* packages to induce decision-trees have been compared in modelling the presence–absence of the invasive Iberian gudgeon. The optimal parameters and the variable subset were simultaneously sought with a GA obtaining accuracy (TSS) similar to studies that used decision-trees to classify the presence or absence of other aquatic organisms (*e.g.*, Parasiewicz et al., 2012; Sharma et al., 2009). Results indicated no marked differences on the predictive capability of any of the techniques and the accuracy obtained for the four decision-trees was marginally superior to other comparable examples

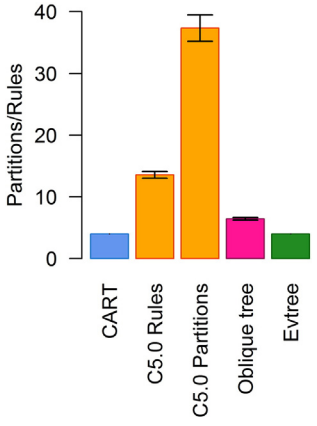

Fig. 3. Complexity of the decision-trees as the overall number of partitions (*i.e.* sum of the number of times a variable was used to split the input space) during the $3\times3_{cross-validation}$ (nine models). For C5.0, the number of rules as well as the overall number of partitions are depicted.

that used logistic regression, although accuracy depends on the interaction between the technique and the training dataset (Parasiewicz et al., 2012; Vezza et al., 2012). However, with respect to model performance, random forests (*i.e.* the ensemble counterpart of CARTs) could outperform the four methods we applied in this study (as was shown in Vezza et al., 2015). The random forests technique typically achieves higher performance than the others (*e.g.*, Fukuda et al., 2013; Vezza et al., 2014) although it always requires indirect methods to be scrutinised (Friedman, 2001), which may discourage its use when high interpretability is desired.

Focusing on the results obtained through the cross-validation and the optimal single model, C5.0 performed slightly better than the other three approaches. Therefore, in spite of having selected a larger input subset (seven variables instead of only two), the larger amount of optimised parameters and the possibility of converting the decision-tree into a rule-base model allowed C5.0 to sustain good generalisation. This suggests C5.0 to be an appealing technique, especially when accuracy must prevail. Further, while we disabled the option of boosting and ensemble of decision-trees, such options may potentially result in accuracy similar to random forests but it will be always at the expense of model interpretability. Moreover, the use of rules, especially when they become as complex as the ones rendered in this case, drives out the interpretability, which is one of the appealing characteristics of decision-trees. Therefore depending on the objectives of the study C5.0 should not automatically relegate the other alternatives.

Although the oblique tree used oblique splits, it did not signify a marked improvement in accuracy and the modelled patterns of habitat selection were practically the same as those obtained with CART. These results agreed with other studies on oblique tree induction where they achieved similar or less accurate results than with the axis-parallel approaches (Heath et al., 1993; Murthy et al., 1994), although other studies do not allow comparison because they used only oblique decision-tree approaches (Cantú-Paz and Kamath, 2003). Thus, it can be arguably concluded that oblique trees do not render obligatory better results either in terms of accuracy and/or in terms of flexibility to adjust the discriminant surface. However, taking into account that model performance depends on the interaction between the input data and the training algorithm and considering that the lapse of the optimisation was not excessive, a trial with oblique trees would be always worthwhile.

Conversely, evtree was found to be practically the same decision-tree as CART, they only slightly differed in the values of the splits (*e.g.* from 0.035 to 0.04 boulders/m^2), but the lapse in the optimisation rose two orders of magnitude. Further, the authors themselves acknowledged that evtree was not intended to relegate other approaches to develop decision-trees, rather to enable scientist to explore different facets of data structure by testing different data partitions (Grubinger et al., 2014). In order to keep comparison focused on the methods for decision-tree induction, as far as possible, we used the same approach and settings for the GA, but other strategies could be used. The *alpha* parameter is able to eventually control tree complexity and hence the ultimate set of selected variables. Therefore an alternative approach may rely on the modification of *alpha*. Several values could be tested, keeping a sufficiently large population and number of generations, and finally selecting the value of *alpha* that maximises accuracy and minimises the size of the variable subset (*e.g.* following Muñoz-Mas et al., 2015).

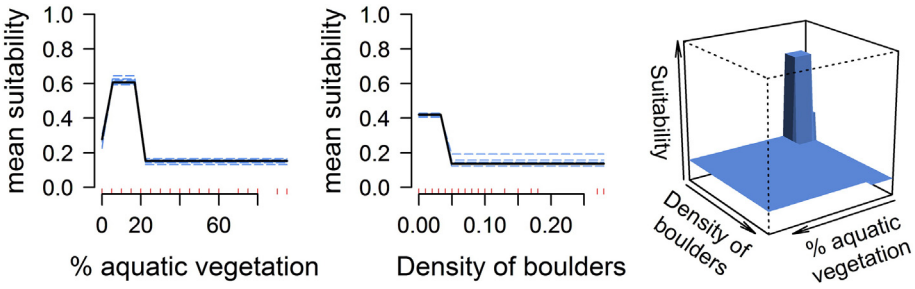

Fig. 4. Partial dependence plots of the classification and regression tree (CART). The solid line depicts those for the ultimate model whereas dashed lines correspond to the cross-validated models. Ticks close to the x-axis depict collected data. Only two variables were selected, therefore on the right side, the 3D plot depicts the variables and the corresponding suitability (probability of presence).

In that case, we would expect a significant reduction in the time of calculus, which would increase interest in the algorithm. Furthermore, unlike *C50* (Kuhn et al., 2015) and Truong's (2009) approach the *evtree* package (Grubinger et al., 2014) is able to optimise regression trees thus it could increase significantly the value of the package.

One common pitfall observed in the four methods has been underprediction (sensitivity < specificity), which was of similar magnitude as the one observed using datasets of similar prevalence (Parasiewicz et al., 2012; Sharma et al., 2009). Such demeanour has been noted as being hardly defensible from an ecological viewpoint (Fukuda, 2013) because, while presences indicate the use of such habitat, absences indicate uncertainty. They may signify unsuitable habitat but they can be also caused by a deficient colonisation or by a low probability of detection (Fukuda, 2013; Muñoz-Mas et al., 2014). It is therefore necessary to at least balance the omission and commission error instead of favouring the majority class (Muñoz-Mas et al., 2014). The prevalence of the dataset was enviable in comparison with other studies (Parasiewicz et al., 2012; Muñoz-Mas et al., 2012). Therefore, it posed no major concern and the only precaution relied on maintaining the original prevalence in every fold. Common approaches to favour overprediction consist of weighting cases (*e.g.* Parasiewicz et al., 2012) or resampling to obtain the desired prevalence (*e.g.* Muñoz-Mas et al., 2012). All the tested packages except oblique.tree allow case weighting, which may discourage its use. Furthermore, the package implemented to develop the oblique tree ensembles (*i.e.* obliqueRF — Menze et al., 2011) suffers similar limitation. Therefore, it can be concluded that the development of oblique decision-trees (single or assembled) in *R* (R Core Team, 2015) are still in an incipient stage and thus further enhancement of these packages with respect to case weighting or resampling is needed. Summing up, C5.0 (*C50*, Kuhn et al., 2015) would be the most appealing technique in accordance with the actual implementation of the other packages.

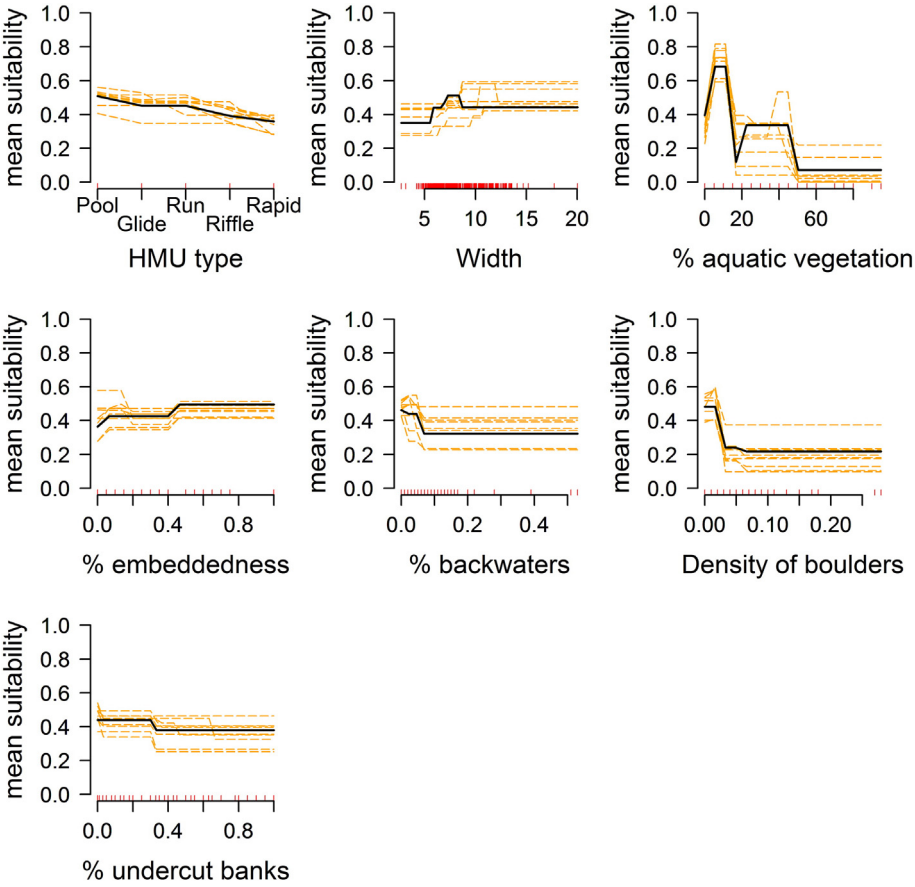

Fig. 5. Partial dependence plots of the C5.0 model tree. The solid line depicts those for the ultimate model whereas dashed lines correspond to the cross-validated models. Ticks close to the x-axis depict collected data.

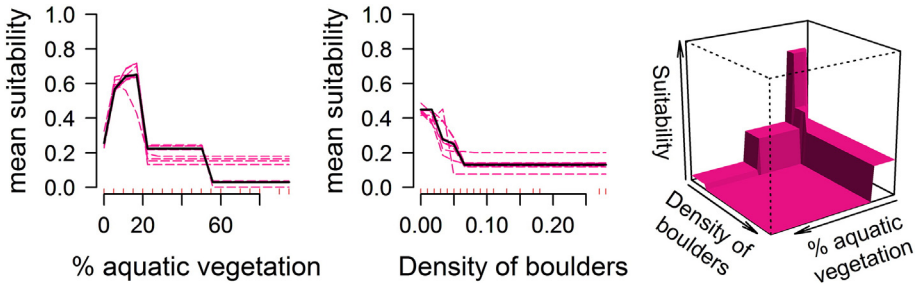

Fig. 6. Partial dependence plots of the oblique tree. The solid line depicts those for the ultimate model whereas dashed lines correspond to the cross-validated models. Ticks close to the x-axis depict collected data. Only two variables were selected, therefore on the right side, the 3D plot depicts the variables and the corresponding suitability (probability of presence).

4.2. Habitat preferences of the Iberian gudgeon

In regard to the habitat preferences of the Iberian gudgeon, the four models indicated % of aquatic vegetation and density of boulders as the most important predictors rendering similar partial dependence plots, whereas C5.0 increased the number of selected predictors up to seven. Though *prima facie* the partial dependence plots for % of aquatic vegetation and density of boulders would be uninformative for management purposes, they provided evidence of the actual distribution and habitat preferences. Firstly, both variables describe not only the study sites, but also some HMU characteristics. On the one hand, at C1 the Cabriel River crosses a narrow canyon with a high gradient, which turns C1 in the river segment with the highest heterogeneity (it presented the largest number of HMUs with the smaller HMU areas). As a consequence, the numerous boulders fallen in the river channel raised the values of density of boulders for that segment, which presented low Iberian gudgeon occurrence. Although C1 presents several pools, typically colonised by the Iberian gudgeon, it is characterised by riffle and rapid type HMUs with relatively high flow velocity, which in accordance with the partial dependence plots for C5.0 are unfavourable for its settlement. On the other hand, rapids are typically characterised by a small area and prominent boulders, thus the density of boulders was evaluated to be high in the remaining river segments. Further, the Iberian gudgeon has been demonstrated to prefer lowland river segments (Comesaña and Ayres, 2009; Ilhéu et al., 2014; Ribeiro et al., 2009), which is concordant with the partial dependence plot for width though the ultimate model suggested a decrease for the larger width several folds suggested monotonic increments. Nevertheless, the Iberian gudgeon proved adaptive in their habitat requirements. For instance, in the Segura River it colonised the river basin through the water transfer canal that disembogues in the upstream part of the river (Martinez-Morales et al. 2010). Thereby the greater part of the population occurs in the headwater rather than in the low part of the river. Apparently, in the Cabriel River the colonisation took place between C3 and C4 (there is also another invasive species in that segment, the Iberian nase) and thus the distribution would be more consonant with the habitat preferences described in the literature (Comesaña and Ayres, 2009; Ilhéu et al., 2014; Ribeiro et al., 2009). In regards to the % of aquatic vegetation, the Iberian gudgeon selected preferably low values but it avoided the HMU without aquatic vegetation. The riverbed in C1 and C2 was largely covered by aquatic vegetation, the former by aquatic liverworts and mosses (*e.g. Fossombronia* sp.) whereas the latter was covered by a thick layer of tangled and mineralised macrophytes (*e.g. Chara* sp.). Such kind of vegetation did not provided cover whereas Iberian gudgeon was observed taking shelter in the vicinity of clumps of *Potamogeton* sp., then the preference for small to intermediate % of aquatic vegetation described in the partial dependence plots. That pattern on habitat selection corroborates previous studies that found larger densities in those sites with intermediate abundances of aquatic vegetation followed by those with large abundances and lastly by those sites without vegetation (Lobon-Cervia et al., 1991). Finally, C4, which is the lowermost stretch, presents typically the highest flow, which combined with the numerous bends on the river channel, favoured scouring (% of undercut banks). Scouring impeded the establishment of vegetation in the shores (% reeds) – which we expected to be selected – but also favoured the % of backwaters (which was also relevant in C1) thus allowed us to deduce that turbulence and lack of cover would be avoided by the Iberian gudgeon. Such a conclusion would be supported by the preference described in the partial dependence plot for % of embeddedness. Therefore the optimal habitat for the Iberian gudgeon would be those pool-type HMUs with a certain amount of aquatic vegetation (providing cover) and depositional areas (most probably in near vegetated shores) that typically occur in lower river segments. Such overall description largely matches the correlation between aquatic vegetation, low flow velocity and presence of bedrock described by Leunda et al. (2012).

In accordance with such general description of the habitat preferences, the widespread of the species (Comesaña and Ayres, 2009; Ilhéu et al., 2014; Ribeiro et al., 2009) may be a reflex of the extensive homogenisation on the habitat conditions occurred in the Iberian Peninsula due to the intense river regulation, channelisation and artificial

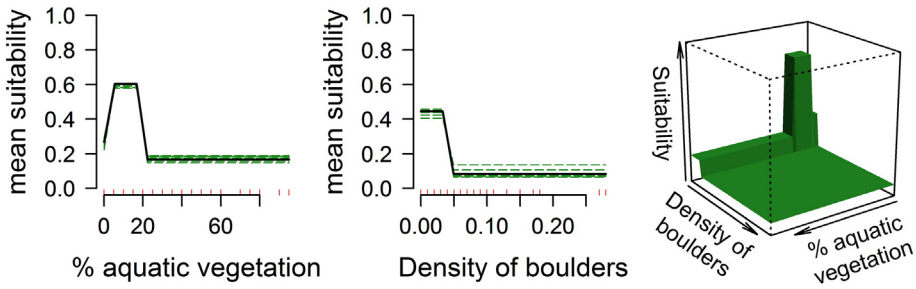

Fig. 7. Partial dependence plots for the evolutionary tree (evtree). The solid line depicts those for the ultimate model whereas dashed lines correspond to the cross-validated models. Ticks close to the x-axis depict collected data. Only two variables were selected, therefore on the right side, the 3D plot depicts the variables and the corresponding suitability (probability of presence).

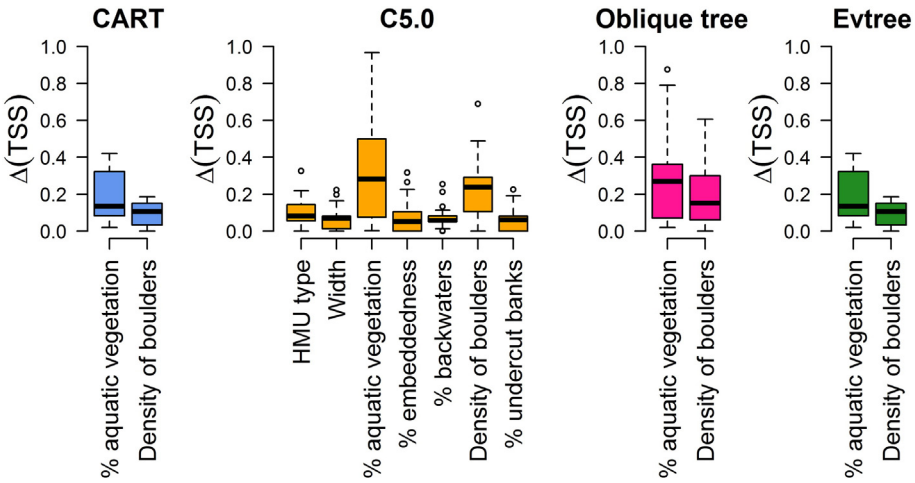

Fig. 8. Variable importance computed by the four decision-tree approaches. The greatest importance corresponds to the variable showing the highest variability and magnitude (Δ*TSS*) whereas the least important presents the smallest.

impoundment (García de Jalón, 1987). Therefore, removing obsolete weirs and dams and thus reducing the total area of pool-like HMUs is likely to reduce the species proliferation (Olaya-Marín et al., 2012) whereas incrementing pulse flow may have also benefits tearing off the clumps of aquatic vegetation and disturbing the depositional areas (% of embeddedness) where the species takes shelter (Tena et al., 2013). Finally, geomorphological restoration and channel re-meandering (*i.e.* increasing the % of backwaters) should discourage the settlement of the species. Fortunately, these particular management actions are emphasised in the European Common Implementation Strategy (CIS) guidance on Environmental Flows (EU-CIS guidance No. 31) as possible hydro-morphological restoration measures for rivers (European Commission, 2015).

5. Conclusions

None of the new packages, *oblique.tree* (Truong, 2013) and *evtree* (Grubinger et al., 2014), outperformed the C5.0 algorithm implemented in the package *C50* (Kuhn et al., 2015). Rather they rendered practically the same decision-trees as the CART developed with the package *tree* (Ripley, 2015), although they were fully interpretable in comparison with C5.0 that was largely complex. We conclude that oblique trees do not necessarily represent an improvement in accuracy in spite of the flexible model structure because it resulted in a similar discriminant surface (the same variables were selected). Further it proved affected by prevalence, a drawback that cannot be easily addressed because it does not include the possibility of weighting the observations. Although the lapse of optimisation of *evtree* was significantly longer, other approaches can be followed, which most probably will reduce it, but maintaining the quality of the ultimate decision-tree. Further it allows the development of regression trees which may be interesting for abundance-related modelling tasks. In unspoilt rivers where physical–chemical predictors do not condition the distribution of the Iberian gudgeon, the optimal habitat for the Iberian gudgeon would be large pools in lowland river segments with depositional areas and vegetated cover present, which typically appeared in the form of isolated and scattered macrophytes clumps (low to intermediate % of aquatic vegetation). Thereby it avoided habitats characterised high flow velocity, which increased bank erosion and limited cover availability. In accordance with these results, the spread of the Iberian gudgeon may have been favoured by river regulation and artificial river impoundment.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ecoinf.2016.04.011.

Acknowledgments

The study has been partially funded by the national Research project IMPADAPT (CGL2013-48424-C2-1-R) with MINECO (Spanish Ministry of Economy) and Feder funds and by the Confederación Hidrográfica del Júcar (Spanish Ministry of Agriculture, Food and Environment). This study was also supported in part by the University Research Administration Center of the Tokyo University of Agriculture and Technology. Finally, we are grateful to the colleagues who worked in the field data collection, especially Juan Diego Alcaráz-Henández, Rui M. S. Costa and Aina Hernández.

References

Adriaenssens, V., De Baets, B., Goethals, P.L.M., De Pauw, N., 2004. Fuzzy rule-based models for decision support in ecosystem management. Sci. Total Environ. 319 (1–3), 1–12. http://dx.doi.org/10.1016/S0048-9697(03)00433–9.

Alcaraz, C., Carmona-Catot, G., Risueño, P., Perea, S., Pérez, C., Doadrio, I., et al., 2014. Assessing population status of *Parachondrostoma arrigonis* (Steindachner, 1866), threats and conservation perspectives. Environ. Biol. Fish 98 (1), 443–455. http://dx.doi.org/10.1007/s10641-014-0274-3.

Almeida, D., Grossman, G.D., 2012. Utility of direct observational methods for assessing competitive interactions between non-native and native freshwater fishes. Fish. Manag. Ecol. 19 (2), 157–166. http://dx.doi.org/10.1111/j.1365-2400.2012.00847.x.

Almeida, D., Ribeiro, F., Leunda, P.M., Vilizzi, L., Copp, G.H., 2013. Effectiveness of FISK, an invasiveness screening tool for non-native freshwater fishes, to perform risk identification assessments in the Iberian Peninsula. Risk Anal. 33 (8), 1404–1413. http://dx.doi.org/10.1111/risa.12050.

Aparicio, E., Carmona-Catot, G., Kottelat, M., Perea, S., Doadrio, I., 2013. Identification of *Gobio* populations in the northeastern Iberian Peninsula: first record of the non-native Languedoc gudgeon *Gobio occitaniae* (Teleostei, Cyprinidae). BioInvasions Rec. 2 (2), 163–166. http://dx.doi.org/10.3391/bir.2013.2.2.13.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. http://dx.doi.org/10.1214/09-SS054.

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.

Bain, M.B., Finn, J.T., Booke, H.E., 1985. A quantitative method for sampling riverine microhabitats by electrofishing. N. Am. J. Fish Manag. 5 (3), 489–493. http://dx.doi.org/10.1577/1548-8659(1985)52.0.CO;2.

Baxter, K., Shortis, M., 2002. Identifying fish habitats: the use of spatially explicit habitat modeling and prediction in marine research. SIRC 2002 — The 14th Annual Colloquium of the Spatial Information Research Centre, Dunedin (New Zealand), pp. 121–130.

Borra, S., Di Ciaccio, A., 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comput. Stat. Data Anal. 54 (12), 2976–2989. http://dx.doi.org/10.1016/j.csda.2010.03.004.

Bovee, K.D., Lamb, B.L., Bartholow, J.M., Stalnaker, C.B., Taylor, J., Henriksen, J., 1998. Stream Habitat Analysis Using the Instream Flow Incremental Methodology Geological Survey — Information and Technology Report 1998-0004. Fort Collins, CO (USA), p. 130.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Chapman & Hall (368 pp.).

Cantú-Paz, E., Kamath, C., 2003. Inducing oblique decision trees with evolutionary algorithms. IEEE Trans. Evol. Comput. 7 (1), 54–68. http://dx.doi.org/10.1109/TEVC.2002.806857.

Clavero, M., 2011. Assessing the risk of freshwater fish introductions into the Iberian Peninsula. Freshw. Biol. 56 (10), 2145–2155. http://dx.doi.org/10.1111/j.1365-2427.2011.02642.x.

Clavero, M., García-Berthou, E., 2006. Homogenization dynamics and introduction routes of invasive freshwater fish in the Iberian Peninsula. Ecol. Appl. 16 (6), 2313–2324. http://dx.doi.org/10.1890/1051-0761(2006)016[2313:HDAIRO]2.0.CO;2.

Comesaña, J., Ayres, C., 2009. New data on the distribution of pumpkinseed *Lepomis gibbosus* and largemouth bass *Micropterus salmoides*, and of non endemic Iberian gudgeon *Gobio lozanoi* in the Galicia region (NW Spain). Aquat. Invasions 4 (2), 425–427. http://dx.doi.org/10.3391/ai.2009.4.2.25.

Confederación Hidrográfica del Júcar, 2005. Informe para la Comisión Europea sobre los artículos 5 y 6 de la Directiva Marco del Agua. Valencia (Spain). p. 528.

Costa, R.M.S., Martínez-Capel, F., Muñoz-Mas, R., Alcaraz-Hernández, J.D., Garófano-Gómez, V., 2012. Habitat suitability modelling at mesohabitat scale and effects of dam operation on the endangered Júcar nase, *Parachondrostoma arrigonis* (Cabriel River, Spain). River Res. Appl. 28 (6), 740–752. http://dx.doi.org/10.1002/rra.1598.

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., et al., 2007. Random forests for classification in ecology. Ecology 88 (11), 2783–2792. http://dx.doi.org/10.1890/07-0539.1.

Czajkowski, M., Kretowski, M., 2013. Global induction of oblique model trees: an evolutionary approach. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (Eds.), Artificial Intelligence and Soft Computing. Springer, Berlin Heidelberg, Berlin (Germany), pp. 1–11.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2006. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecol. Model. 195 (1–2), 20–29. http://dx.doi.org/10.1016/j.ecolmodel.2005.11.005.

Doadrio, I., 2002. Atlas y Libro Rojo de los Peces Continentales de España. Dirección General de Conservación de la. Naturaleza, Museo Nacional de Ciencias Naturales, Madrid (Spain) (376 pp.).

Doadrio, I., Madeira, M.J., 2004. A new species of the genus *Gobio* Cuvier, 1816 (Actynopterigii, Cyprinidae) from the Iberian Peninsula and southwestern France. Graellsia 60 (1), 107–116. http://dx.doi.org/10.3989/graellsia.2004.v60.i1.

Esposito, F., Malerba, D., Semeraro, G., 1997. A comparative analysis of methods for pruning decision trees. IEEE Trans. Pattern Anal. Mach. Intell. 19 (5), 476–491. http://dx.doi.org/10.1109/34.589207.

European Commision, 2015. Ecological Flows in the Implementation of the Water Framework Directive — Guidance Document No. 3. Luxembourg (Luxembourg). p. 108.

Fogel, D.B., 1994. Introduction to simulated evolutionary optimization. IEEE Trans. Neural. Netw. 5 (1), 3–14. http://dx.doi.org/10.1109/72.265956.

Fox, J., 2010. Polycor: Polychoric and Polyserial Correlations. R Package Version 0.7–8.

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., 2013. Effects of data prevalence on species distribution modelling using a genetic Takagi–Sugeno fuzzy system. IEEE International Workshop on Genetic and Evolutionary Fuzzy Systems (GEFS), Singapore, pp. 21–27.

Fukuda, S., De Baets, B., 2012. A short review on the application of computational intelligence and machine learning in the bioenvironmental sciences. Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), Kobe (Japan), pp. 106–110.

Fukuda, S., Mouton, A.M., De Baets, B., 2011a. 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., Nakajima, J., De Baets, B., Waegeman, W., Mukai, T., Mouton, A.M., Onikura, N., 2011b. A discussion on the accuracy-complexity relationship in modelling fish habitat preference using genetic Takagi–Sugeno fuzzy systems. IEEE 5th International Workshop on Genetic and Evolutionary Fuzzy Systems (GEFS), Paris (France), pp. 81–86.

Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J., 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. Model. Softw. 47, 1–6. http://dx.doi.org/10.1016/j.envsoft.2013.04.005.

Fukuda, S., Tanakura, T., Hiramatsu, K., Harada, M., 2014. Assessment of spatial habitat heterogeneity by coupling data-driven habitat suitability models with a 2D hydrodynamic model in small-scale streams. Ecol. Inf. http://dx.doi.org/10.1016/j.ecoinf.2014.10.003.

Galathiya, A., Ganatra, A., Bhensdadia, C., 2012. Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning. Int. J. Comp. Sci. Inf. Technol. 3 (2), 3427–3431.

García de Jalón, D., Schmidt, G., 1995. Manual práctico Para la gestión Sostenible de la Pesca Fluvial (Madrid, (Spain)) IEEE, p. 169.

García de Jalón, D.G., 1987. River regulation in Spain. Regul. Rivers Res. Manag. 1 (4), 343–348. http://dx.doi.org/10.1002/rrr.3450010406.

Garofalakis, M., Hyun, D., Rastogi, R., Shim, K., 2000. Efficient Algorithms for Constructing Decision Trees with Constraints. KDD '00 — Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA (USA), pp. 335–339.

German, D.M., Adams, B., Hassan, A.E., 2013. The evolution of the R software ecosystem. 17th European Conference on Software Maintenance and Reengineering (CSMR), Genova (Italy), pp. 243–252.

Gozlan, R.E., Britton, J.R., Cowx, I., Copp, G.H., 2010. Current knowledge on non-native freshwater fish introductions. J. Fish Biol. 76 (4), 751–786. http://dx.doi.org/10.1111/j.1095-8649.2010.02566.x.

Grubinger, T., Zeileis, A., Pfeiffer, K., 2014. Evtree: evolutionary learning of globally optimal classification and regression trees in R. J. Stat. Softw. 61 (1), 1–29.

Heath, D., Kasif, S., Salzberg, S., 1993. Induction of oblique decision trees. J. Artif. Intell. Res. 2 (2), 1002–1007.

Heggenes, J., Brabrand, Å., Saltveit, S., 1990. Comparison of three methods for studies of stream habitat use by young brown trout and Atlantic salmon. Trans. Am. Fish. Soc. 119 (1), 101–111. http://dx.doi.org/10.1577/1548-8659(1990)1192.3.co;2.

Holland, J.H., 1992. Genetic algorithms. Sci. Am. 267 (1), 66–72.

Huang, C., Wang, C., 2006. A GA-based feature selection and parameters optimization for support vector machines. Expert Syst. Appl. 31 (2), 231–240. http://dx.doi.org/10.1016/j.eswa.2005.09.024.

Ilhéu, M., Matono, P., Bernardo, J.M., 2014. Invasibility of Mediterranean-climate rivers by non-native fish: the importance of environmental drivers and human pressures. PLoS One 9 (11). http://dx.doi.org/10.1371/journal.pone.0109694.

Instituto Nacional de Estadística (INE), 2013. Proyección de la Población de España a Corto Plazo (2013–2023) (Madrid (Spain)) p. 96.

Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J., 2004. Filter versus wrapper gene selection approaches in DNA microarray domains. Artif. Intell. Med. 31 (2), 91–103. http://dx.doi.org/10.1016/j.artmed.2004.01.007.

Kass, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. J. R. Stat. Soc.: Ser. C: Appl. Stat. 29 (2), 119–127. http://dx.doi.org/10.2307/2986296.

Kemp, S.J., Zaradic, P., 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.

Kuhn, M., Weston, S., Coulter, N., Culp, M., 2015. C50: C5.0 Decision Trees and Rule-based Models (C Code for C5.0 by R. Quinlan). R Package Version 0.1.0-24.

Kwon, Y.-S., Bae, M.-J., Hwang, S.-J., Kim, S.-H., Park, Y.-S., 2015. Predicting potential impacts of climate change on freshwater fish in Korea. Ecol. Inform. 29, 156–165. http://dx.doi.org/10.1016/j.ecoinf.2014.10.002.

Leclere, J., Oberdorff, T., Belliard, J., 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. http://dx.doi.org/10.1016/j.ecoinf.2011.05.001.

Leunda, P.M., 2010. Impacts of non-native fishes on Iberian freshwater ichthyofauna: current knowledge and gaps. Aquat. Invasions 5 (3), 239–262. http://dx.doi.org/10.3391/ai.2010.5.3.03.

Leunda, P.M., Sistiaga, M., Oscoz, J., Miranda, R., 2012. Ichthyofauna of a near-natural Pyrenean river: spatio-temporal variability and reach-scale habitat. Environ. Eng. Manag. J. 11 (6), 1111–1124.

Liaw, A., Wiener, M., 2002. Classification and regression by *randomForest*. R News 3 (2), 18–22.

Lin, Y., Lin, W., Wu, W., 2015. Uncertainty in various habitat suitability models and its impact on habitat suitability estimates for fish. Water 7 (8), 4088–4107. http://dx.doi.org/10.3390/w7084088.

Lobon-Cervia, J., Montañes, C., De Sostoa, A., 1991. Influence of environment upon the life history of gudgeon, *Gobio gobio* (L.): a recent and successful colonizer of the Iberian Peninsula. J. Fish Biol. 39 (3), 285–300. http://dx.doi.org/10.1111/j.1095-8649.1991.tb04363.x.

Martínez-Morales, I., Oliva-Paterna, F.J., Verdiell-Cubedo, D., Torralva, M., 2010. Inventario y estado de conservación de la fauna piscícola en la cuenca alta del río Segura (SE Península Ibérica). An. Biol. 32, 47–58.

Mebane Jr., W.R., Sekhon, J.S., 2011. Genetic optimization using derivatives: the rgenoud package for R. J. Stat. Softw. 42 (11), 1–26.

Menze, B.H., Kelm, B.M., Splitthoff, D.N., Koethe, U., Hamprecht, F.A., 2011. On Oblique Random Forests. Lect. Notes Comput. Sci. 6912 LNAI, PART 2453-469 http://dx.doi.org/10.1007/978-3-642-23783-6_29.

Miñano, P.A., García-Mellado, A., Oliva-Paterna, F.J., Torralva, M., 2003. Edad, crecimiento y reproducción de *Gobio gobio* L. (Pisces, Cyprinidae) en un tramo regulado del río Segura (SE España). Anim. Biodivers. Conserv. 26 (1), 67–76.

Mouton, A.M., Alcaraz-Hernández, J.D., De Baets, B., Goethals, P.L.M., Martínez-Capel, F., 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. Environ. Model. Softw. 26 (5), 615–622. http://dx.doi.org/10.1016/j.envsoft.2010.12.001.

Muñoz-Mas, R., Martínez-Capel, F., Alcaraz-Hernández, J.D., Mouton, A.M., 2015. Can multilayer perceptron ensembles model the ecological niche of freshwater fish species? Ecol. Model. 309–310 (0), 72–81. http://dx.doi.org/10.1016/j.ecolmodel.2015.04.025.

Muñoz-Mas, R., Martínez-Capel, F., Garófano-Gómez, V., Mouton, A.M., 2014. Application of probabilistic neural networks to microhabitat suitability modelling for adult brown trout (*Salmo trutta* L.) in Iberian rivers. Environ. Model. Softw. 59 (0), 30–43. http://dx.doi.org/10.1016/j.envsoft.2014.05.003.

Muñoz-Mas, R., Martínez-Capel, F., Schneider, M., 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.

Murthy, S.K., Kasif, S., Salzberg, S., 1994. A system for induction of oblique decision trees. J. Artif. Intell. Res. 2 (1), 1–32. http://dx.doi.org/10.1613/jair.63.

Olaya-Marín, E.J., Martínez-Capel, F., Soares Costa, R.M., 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. http://dx.doi.org/10.1016/j.scitotenv.2012.07.093.

Olden, J.D., 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., 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.

Oscoz, J., Leunda, P.M., Miranda, R., Escala, M.C., 2006. Summer feeding relationships of the co-occurring *Phoxinus phoxinus* and *Gobio lozanoi* (Cyprinidae) in an Iberian river. Folia Zool. 55 (4), 418–432.

Pandey, H.M., Chaudhary, A., Mehrotra, D., 2014. A comparative review of approaches to prevent premature convergence in GA. Appl. Soft Comput. J. 24, 1047–1077. http://dx.doi.org/10.1016/j.asoc.2014.08.025.

Parasiewicz, P., Castelli, E., Rogers, J.N., Plunkett, E., 2012. Multiplex modeling of physical habitat for endangered freshwater mussels. Ecol. Model. 228, 66–75. http://dx.doi.org/10.1016/j.ecolmodel.2011.12.023.

Quinlan, J.R., 1992. C4. 5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., San Mateo, CA (USA) (302 pp.).

R Core Team, 2015. R: A Language and Environment for Statistical Computing. Version 3.2.1.

Reyjol, Y., Hugueny, B., Pont, D., Bianco, P.G., Beier, U., Caiola, N., et al., 2007. Patterns in species richness and endemism of European freshwater fish. Glob. Ecol. Biogeogr. 16 (1), 65–75. http://dx.doi.org/10.1111/j.1466-8238.2006.00264.x.

Ribeiro, F., Collares-Pereira, M.J., Moyle, P.B., 2009. Non-native fish in the fresh waters of Portugal, Azores and Madeira Islands: a growing threat to aquatic biodiversity. Fish. Manag. Ecol. 16 (4), 255–264. http://dx.doi.org/10.1111/j.1365-2400.2009.00659.x.

Ribeiro, F., Elvira, B., Collares-Pereira, M.J., Moyle, P.B., 2008. Life-history traits of non-native fishes in Iberian watersheds across several invasion stages: a first approach. Biol. Invasions 10 (1), 89–102. http://dx.doi.org/10.1007/s10530-007-9112-2.

Ripley, B., 2015. Tree: Classification and Regression Trees. R Package Version 1.0-36.

Sadeghi, R., Zarkami, R., Sabetraftar, K., Van Damme, P., 2013. Application of genetic algorithm and greedy stepwise to select input variables in classification tree models for the prediction of habitat requirements of *Azolla filiculoides* (Lam.) in Anzali wetland, Iran. Ecol. Model. 251, 44–53. http://dx.doi.org/10.1016/j.ecolmodel.2012.12.010.

Schill, D.J., Griffith, J.S., 1984. Use of underwater observations to estimate cutthroat trout abundance in the Yellowstone River. N. Am. J. Fish. Manag. 4 (4), 479–487. http://dx.doi.org/10.1577/1548-8659(1984)42.0.CO;2.

Sharma, S., Herborg, L., Therriault, T.W., 2009. Predicting introduction, establishment and potential impacts of smallmouth bass. Divers. Distrib. 15 (5), 831–840. http://dx.doi.org/10.1111/j.1472-4642.2009.00585.x.

Stein, G., Chen, B., Wu, A.S., Hua, K.A., 2005. Decision tree classifier for network intrusion detection with GA-based feature selection. 43rd Annual Association for Computing Machinery Southeast Conference (ACMSE '05), Kennesaw, GA (USA, pp. 2136–2141.

Tena, A., Ksiaszek, L., Vericat, D., Batalla, R.J., 2013. Assessing the geomorphic effects of a flushing flow in a large regulated river. River Res. Appl. 29 (7), 876–890. http://dx.doi.org/10.1002/rra.2572.

Therneau, T., Atkinson, B., Ripley, B., 2015. rpart: Recursive Partitioning and Regression Trees. R package Version 4.1-9.

Truong, A.K.Y., 2009. Fast Growing and Interpretable Oblique Trees via Logistic Regression Models. University of Oxford, Oxford (UK), p. 123.

Truong, A.K.Y., 2013. Oblique.Tree: Oblique Trees for Classification Data. R Package Version 1.1.1.

Vezza, P., Muñoz-Mas, R., Martinez-Capel, F., 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.

Vezza, P., Parasiewicz, P., Calles, O., Spairani, M., Comoglio, C., 2014. Modelling habitat requirements of bullhead (*Cottus gobio*) in alpine streams. Aquat. Sci. 76 (1), 1–15. http://dx.doi.org/10.1007/s00027-013-0306-7.

Vezza, P., Parasiewicz, P., Rosso, M., Comoglio, C., 2012. Defining minimum environmental flows at regional scale: application of mesoscale habitat models and catchments classification. River Res. Appl. 28 (6), 717–730. http://dx.doi.org/10.1002/rra.1571.

Wilkes, M.A., Maddock, I., Link, O., Habit, E., 2015. A community-level, (mesoscale analysis of fish assemblage structure in shoreline habitats of a large river using multivariate regression trees. River Res. Appl. http://dx.doi.org/10.1002/rra.2879 (n/a–n/a).

Wu, X., Kumar, V., Quinlan, R.J., Ghosh, J., Yang, Q., Motoda, H., et al., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. Syst. 14 (1), 1–37. http://dx.doi.org/10.1007/s10115-007-0114-2.